



# **Comparative Evaluation of Machine Learning Models for Reliable Breast Cancer Prediction**

**By**

**Sneha Devjibhai Babariya**

**Supervisor: Mark Huntly**

**Project submitted as part of the requirements for the award of  
M.Sc. Data and Artificial intelligence**

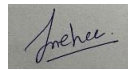
**November 2025**

## Appendix 1: DECLARATION

I, Sneha Devjibhai babariya declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed :

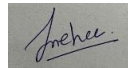


Date : 12 November 2025

## FORM OF CONSENT

**I, Sneha Devjibhai Babariya**, hereby consent that my Project, submitted in candidature for the M.Sc. Data and Artificial intelligence, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed:



Date: 12 November 2025

## Abstract

The purpose of this study is to investigate on a comparative assessment of machine learning-based predictors of reliable breast cancer by ultrasound image data. It was done with the purpose to find the most accurate, interpretable and computationally efficient model to aid with early diagnosis and clinical decision-making. The study assumed a positivist philosophy and extensive use of deductive, quantitative designs, using supervised machine learning methods such as Logistic Regression, Decision Tree, Random Forest, Support Vector, whilst K-Nearest Neighbour, Artificial Neural Network, and deep learning networks, including a YOLO-based Convolutional Neural Network (CNN), ResNet50, and VGG16 and a proposed Hybrid CNN + handcrafted feature model. A dataset of Breast Ultrasound pictures (BUSI) on Kaggle was processed by means of contrast-enhancement (CLAHE) technology together with the area-of-interest extraction, colour and normalisation, as well as augmentation and evaluation to enhance content robustness in the models. The accuracy, precision, recall, F1-score and Precision–Recall (PR) were used to evaluate models. It was found that the greatest accuracy (87.18), macro F1-score (0.8689), which outperforms ResNet50 (81.41%) and VGG16 (81.41) and achieved comparable performance with the Hybrid CNN model (83.33). Incremental validation was done to verify the statistical significance of these differences ( $p < 0.05$ ), and computational validation was done to show how the CNN is efficient in a small number of parameters and can also infer within a short time. The results of the study point to the possibility of lightweight CNNs and hybrid feature fusion models being applied in medical practice and primarily in the resource-restricted setting, as they offer the best balance between the factors of diagnostic precision, explainability, and scalability. The research concludes that tailor-made CNN and hybrid deep learning models can promote the process of detecting breast cancer at an early stage and recommends their implementation in the real-time diagnostics tool. The recommended future work is to use bigger and multi-institutional datasets, explainable AI integration, and federated learning to expand the application scope of clinical use.

**Keywords:** Breast Cancer Prediction, Machine Learning, Deep Learning, Convolutional Neural Network (CNN), Ultrasound Imaging, Comparative Analysis, YOLO Model, Random Forest, Support Vector Machine (SVM), Clinical Decision Support

## Table of Contents

|                                                                      |    |
|----------------------------------------------------------------------|----|
| Appendix 1: DECLARATION .....                                        | 2  |
| Table of Contents .....                                              | 4  |
| List of Figures .....                                                | 7  |
| List of Tables .....                                                 | 8  |
| Chapter 1: Introduction .....                                        | 9  |
| 1.1 Introduction .....                                               | 9  |
| 1.2 Aim and Objectives .....                                         | 10 |
| 1.3 Research Question .....                                          | 10 |
| 1.4 Background and Rationale .....                                   | 10 |
| 1.4.1 Background .....                                               | 10 |
| 1.4.2 Rationale .....                                                | 11 |
| 1.5 Research Motivation and Scope .....                              | 12 |
| Chapter 2: Literature Review .....                                   | 13 |
| 2.1 Introduction .....                                               | 13 |
| 2.2 Role of Machine Learning and AI in Cancer Prediction .....       | 13 |
| 2.3 Machine Learning Models for Cancer Prediction .....              | 14 |
| 2.3.1 Support Vector Machine (SVM) .....                             | 15 |
| 2.3.2 Random Forest (RF) .....                                       | 16 |
| 2.3.3 Logistic Regression (LR) .....                                 | 16 |
| 2.3.4 Decision Tree (DT) .....                                       | 17 |
| 2.3.5 K-Nearest Neighbour (KNN) .....                                | 18 |
| 2.3.6 Artificial Neural Networks (ANN) .....                         | 19 |
| 2.4 Preprocessing and Feature Engineering in Cancer Prediction ..... | 19 |
| 2.4.1 Preprocessing Techniques .....                                 | 19 |

|                                                                   |    |
|-------------------------------------------------------------------|----|
| 2.4.2 Feature Selection and Representation .....                  | 20 |
| 2.5 Deep Learning and Hybrid Approaches .....                     | 20 |
| 2.6 Applications Beyond Diagnosis.....                            | 21 |
| 2.6.1 Prognosis and Treatment Planning .....                      | 21 |
| 2.6.2 Healthcare Economics and Resource Management.....           | 21 |
| 2.7 Federated Learning and Data Privacy in Cancer Prediction..... | 21 |
| 2.8 Literature Gap .....                                          | 22 |
| Chapter 3: Methodology .....                                      | 24 |
| 3.1 Introduction.....                                             | 24 |
| 3.2 Research Philosophy, Design and Approach .....                | 24 |
| 3.3 Tools and Techniques .....                                    | 25 |
| 3.3.1 Data Preprocessing.....                                     | 25 |
| 3.3.2 Feature Extraction.....                                     | 25 |
| 3.3.3 Data Splitting .....                                        | 25 |
| 3.3.4 Model Training .....                                        | 26 |
| 3.3.5 Performance Evaluation.....                                 | 26 |
| 3.4 Data Collection .....                                         | 27 |
| 3.5 Used Libraries and Machine Learning Models.....               | 27 |
| 3.6 Theoretical Framework.....                                    | 29 |
| 3.7 Project Plan .....                                            | 31 |
| 3.8 Ethical Consideration.....                                    | 32 |
| Chapter 4: Results Evaluation and Discussion .....                | 34 |
| 4.1 Introduction to Results.....                                  | 34 |
| 4.2 Results and Analysis .....                                    | 34 |
| 4.2.1 Data Pre-processing and Augmentation Results.....           | 34 |

|                                                                 |    |
|-----------------------------------------------------------------|----|
| 4.2.2 Training Dynamics and Convergence Analysis.....           | 36 |
| 4.3 Model Evaluation.....                                       | 37 |
| 4.3.1 Performance Metrics Overview .....                        | 37 |
| 4.3.2 Detailed Analysis of Model Performance.....               | 37 |
| 4.3.3 Class-Specific Performance Analysis .....                 | 42 |
| 4.3.4 Statistical Significance and Clinical Implications .....  | 44 |
| 4.3.5 Model Optimisation and Fine-Tuning Results.....           | 46 |
| 4.4 Discussion .....                                            | 47 |
| 4.4.1 Statistical Significance and Robustness Analysis.....     | 47 |
| 4.4.2 Computational Efficiency and Resource Utilisation.....    | 48 |
| 4.4.3 Clinical Relevance and Diagnostic Implications .....      | 48 |
| 4.5 Limitations and Future Directions .....                     | 49 |
| 4.6 Literature review comparison with study model results ..... | 49 |
| 4.7 Summary of Results.....                                     | 52 |
| Chapter 5: Conclusion and Future work .....                     | 53 |
| 5.1 Conclusion .....                                            | 53 |
| 5.2 Summary of Achievements.....                                | 54 |
| 5.3 Critical Evaluation .....                                   | 55 |
| 5.4 Limitations .....                                           | 56 |
| 5.5 Future work recommendations. ....                           | 57 |
| References.....                                                 | 59 |

## List of Figures

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| Figure 1.1: Breast cancer statistics globally [1].....                            | 9  |
| Figure 2.1: Clinical applications of AI in early cancer diagnosis [12] .....      | 13 |
| Figure 2.2: Representation of a Support Vector Machine (SVM) classifier [3] ..... | 15 |
| Figure 2.3: RF prediction algorithm [24].....                                     | 16 |
| Figure 2.4: Logistic regression classification model [27].....                    | 17 |
| Figure 2.5: Decision Tree (DT) [7].....                                           | 17 |
| Figure 2.6: K-Nearest Neighbour (KNN) [40] .....                                  | 18 |
| Figure 2.7: ANN model architecture [29].....                                      | 19 |
| Figure 3.1: Research Philosophy, Design and Approach .....                        | 24 |
| Figure 3.2: Theoretical Framework .....                                           | 29 |
| Figure 3.3: Theoretical Framework for hybrid CNN Model .....                      | 30 |
| Figure 3.4: Project Plan.....                                                     | 31 |
| Figure 4.1: Class Distribution .....                                              | 34 |
| Figure 4.2: Sample Images .....                                                   | 35 |
| Figure 4.3: DataFrame after Pre-processing .....                                  | 36 |
| Figure 4.4: YOLO-Inspired CNN Model Performance .....                             | 37 |
| Figure 4.5: ResNet50 Transfer Learning Model Performance .....                    | 38 |
| Figure 4.6: VGG16 Transfer Learning Model Performance .....                       | 39 |
| Figure 4.7: Hybrid (YOLO + Handcrafted) Model Performance .....                   | 39 |
| Figure 4.8: Hybrid (VGG16 + Handcrafted) Model Performance .....                  | 40 |
| Figure 4.9: Hybrid (ResNet50 + Handcrafted) Model Performance .....               | 41 |
| Figure 4.7: Confusion Matrix for All Models.....                                  | 43 |
| Figure 4.8: Precision-Recall for all Models .....                                 | 45 |
| Figure 4.9: Hybrid (VGG16+ Handcrafted)Tuning Result .....                        | 46 |

## List of Tables

|                                                                               |    |
|-------------------------------------------------------------------------------|----|
| Table 1.1: Research Objectives.....                                           | 10 |
| Table 1.2: Current Issues in Breast Cancer Prediction .....                   | 12 |
| Table 2.1: Related Works .....                                                | 15 |
| Table 3.1: Applied Techniques in Methodology .....                            | 25 |
| Table 3.2: Method Steps .....                                                 | 27 |
| Table 3.3: Used Libraries.....                                                | 28 |
| Table 3.4: Developed Models .....                                             | 29 |
| Table 4.1: Comparison between literature review and study model results ..... | 52 |

## Chapter 1: Introduction

### 1.1 Introduction

Breast cancer has been considered one of the most frequent cancers in women all over the world, with a high percentage ratio of morbidity and mortality resulting from the health issue. Early diagnosis is crucial for better survival rates, and medical imaging procedures such as mammography and ultrasound are often involved in the diagnosis process. Therefore, manually interpreting medical images is quite time-consuming; it is all too prone to errors and depends on the presence of professionals [1]. To eliminate these difficulties, the possibility of machine learning (ML) models has become widely discussed in regard to their capability to examine various complicated patterns and offer reasonable diagnostic support.

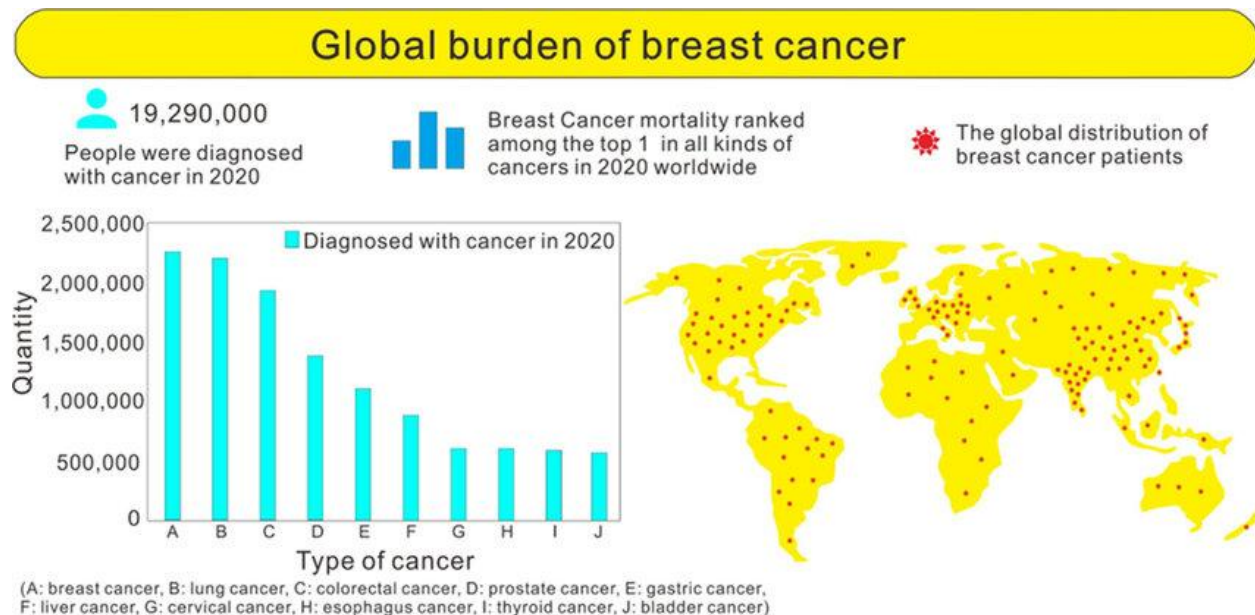


**Figure 1.1: Breast cancer statistics globally [1]**

The project is aimed at the comparative analysis of machine learning models to make breast cancer predictions. The study will identify the best methodology in which various algorithms are applied and evaluated based on the classification of breast ultrasound images to determine whether they were benign, malignant, or even normal. SVM, random forests (RF), logistic regression, K-Nearest Neighbour (KNN), decision trees, and artificial neural networks (ANN) are all tested on important key performance indicators, including accuracy, precision, recall, and F1-score. Studying using a systematic comparison, the project will contribute to the body of continuous research in the field of computational oncology by offering recommendations aimed at creating strong models capable of aiding clinicians in cancer early detection and decision-making.

## 1.2 Aim and Objectives

### Aim

The aim is to evaluate and compare the performance of different machine learning models for reliable breast cancer prediction by identifying the most effective approach that can support early detection and clinical decision-making.

### Objectives

| Objective No.                   | Objective Description                                                                                                                                                                                                               |
|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1. Literature Review            | To review existing research on machine learning techniques applied to breast cancer prediction, highlighting the strengths and limitations of previously used models.                                                               |
| 2. Model Selection              | To identify and select suitable machine learning algorithms (e.g., Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Neural Networks) for comparative analysis in breast cancer prediction by using the Python language. |
| 3. Methodology & Implementation | To design and implement a systematic methodology, including dataset preprocessing, feature selection, model training, and hyperparameter tuning, for fair comparison of the chosen models by using the Python language.             |
| 4. Result Analysis              | To evaluate and compare the performance of the models using appropriate metrics (such as accuracy, precision, recall, F1-score, Precision–Recall (PR) and determine the most reliable model for breast cancer prediction.           |

**Table 1.1: Research Objectives**

## 1.3 Research Question

Which machine learning model among Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and Neural Networks provides the most accurate and reliable prediction of breast cancer, and how do their performances compare across key evaluation metrics?

## 1.4 Background and Rationale

### 1.4.1 Background

The field of artificial intelligence and machine learning has undergone multiple developments that have contributed to issues in the area of healthcare, especially cancer diagnosis. Conventional diagnostic processes, despite their usefulness, have some drawbacks, including errors by human

beings, variation in interpretation, and abuse of high quantities of imagery. It has been shown that machine learning holds a larger amount of accuracy and consistency with respect to the activity of predicting cancer. Selectively, random forests and Support Vector Machines have been documented with accuracy in the classification of breast cancer of more than 90% [3]. Even more powerful predictors are deep-learning-based Convolutional Neural Networks, which extract features automatically based on a complex image [4]. In spite of these developments, there are still unresolved issues on matters pertaining to generalisability, interpretability and clinical adoption. Thus, it is essential to compare alternative algorithms on a comparative basis in order to identify feasible solutions that have to be applied in the practice of breast cancer prediction [8].

#### 1.4.2 Rationale

Although there is strong evidence in the implementation of machine learning across medical imaging, there is only a handful of challenges to the clinical integration of predictive models with regard to breast cancer diagnosis. A significant weakness is the difference in datasets, with limited datasets being used to train many of the models, not representative of diverse populations. This prevents their generalisability in real-life applications. The other is the unequal distribution of datasets; whereby malignant cases are underrepresented as contrasted to the benign ones and lead to biased predictions. Moreover, deep learning models have often been accused of being non-interpretable, even though they are more accurate than alternative methods; clinicians are less likely to be convinced by this aspect [7].

Also, most workplaces place accuracy as the main measure of evaluation without other important evaluation measures like recall or F1-score due to the extreme importance of false negatives in medical environments, where missing a malignant case can lead to delayed diagnosis, increased patient risk, and serious clinical consequences. Computational complexity is also a problem since models such as deep neural networks demand vast resources, which constrain usage in resource-intensive healthcare.

| Issue               | Why This Issue Now                                                                    |
|---------------------|---------------------------------------------------------------------------------------|
| Dataset variability | Models are often trained on limited, homogeneous datasets, reducing generalisability. |
| Class imbalance     | Malignant cases are underrepresented, leading to biased predictions.                  |

|                           |  |                                                                                        |
|---------------------------|--|----------------------------------------------------------------------------------------|
| Interpretability          |  | “Black-box” deep models hinder clinical trust and acceptance.                          |
| Over-reliance on accuracy |  | Lack of multi-metric evaluation ignores false negatives, critical in cancer detection. |
| Computational complexity  |  | High resource requirements restrict use in low-resource clinical environments.         |

**Table 1.2: Current Issues in Breast Cancer Prediction**

Table 1.2 shows the issues and their reason in the chosen sector. Thus, addressing these issues is critical to ensure this by implementing machine learning models. This can prove to be reliable, interpretable, and adaptable for practical healthcare deployment.

### **1.5 Research Motivation and Scope**

The interest in this study is driven by the urgency to detect breast cancer in its early stages with the help of computer-assisted methods of diagnosis that are used to supplement the results of conventional diagnostic procedures [9]. As the level of incidence continues to increase worldwide, an early and precise diagnosis is required to enhance patient outcomes. Current literature has pointed to the promise of machine learning, although there are challenges in inconsistent performance, interpretability, and the use of sample sizes have created a need to conduct a systematic comparative study. The purpose of conducting this research consists of an opportunity to determine the predictive accuracy and clinical relevance of models.

The study is implemented and evaluated by the application and assessment of various machine learning algorithms on the Breast Ultrasound Images Dataset, which defines the scope of the study. The standard metrics to be used to evaluate the comparative performance will be accuracy, precision, recall, and F1-score [10]. Through results analysis in models, the research offers suggestions on reasonable algorithms, which subsequently promote contributions to knowledge in the machine learning field (academically), as well as the application of the algorithms in healthcare (practically).

## Chapter 2: Literature Review

### 2.1 Introduction

This chapter reviews existing research on machine learning approaches for breast cancer prediction. It examines commonly applied models, pre-processing techniques, feature engineering strategies, and performance outcomes. The review is structured into themes to highlight comparative strengths, limitations, and advancements. Through critical analysis of prior studies, this chapter establishes the foundation for the present research and identifies areas requiring further exploration in the development of reliable predictive models for clinical practice.

### 2.2 Role of Machine Learning and AI in Cancer Prediction

Machine learning (ML) and artificial intelligence (AI) have been disruptive technologies in the healthcare sector in this hi-tech world. In this case, these models are implemented, providing the potential to detect the disease at an earlier stage. These would be useful to classify the risk of patients and guide clinicians in how they should treat them. ML models can find non-linear fluctuations of very high-dimensional source material. This is the automatic incorporation of the genomic profiles, clinical records and modalities. This capacity is highly significant in the fight against cancer because it supports detection at very early stages [11].

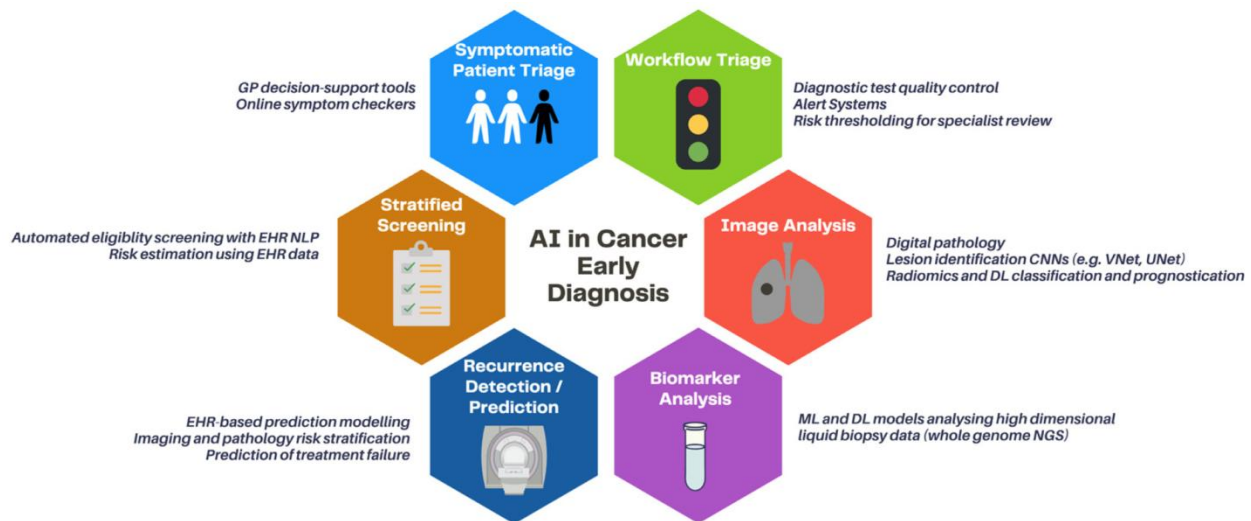


**Figure 2.1: Clinical applications of AI in early cancer diagnosis [12]**

According to many studies, the use of ML applications has been documented to be extremely broad. It finds application in cancer of the breast as well as cervix, cancer of the lung, and also in the pancreas. Such approaches involve the application of classification, clustering and regression models. They assist in empowering healthcare providers to diagnose and estimate the cost of

healthcare, forecasting the length of stay and forecasting hospitalisation as well as surgical problems. Thus, the examples suggest that AI systems are adaptable with regard to the manner they can support in every aspect of cancer management along the healthcare continuum [13]. There is also bias in AI algorithms, particularly when models are trained on datasets that are imbalanced, non-representative, or collected from a single demographic group. Such biases can lead to unequal performance for example, detecting malignant cases less accurately in underrepresented populations which raises concerns about fairness and clinical safety. Recent research therefore stresses the need for transparency and explainability in medical AI systems to render it socially acceptable to clinicians and patients [15].

### 2.3 Machine Learning Models for Cancer Prediction

The predictive modelling in healthcare especially that is done in breast cancer diagnosis and prognosis, is based on machine learning algorithms. These models represents in table 2.1 each offer unique strengths in handling multidimensional and complex medical data, enabling improved clinical decision making. In this section, the discussions will cover the most used models in breast cancer prediction; their mechanisms; their performance; and their relevance.

| Author(s) | Performed Model(s)         | Main Result                                                                |
|-----------|----------------------------|----------------------------------------------------------------------------|
| [14]      | SVM, RF, LR, KNN           | SVM and RF achieved >90% accuracy on the breast cancer dataset.            |
| [5]       | LR, DT, SVM, KNN, RF       | RF and SVM outperformed others; LR and DT showed lower accuracy.           |
| [8].      | ANN, RF, DT                | RF and ANN achieved 92–95% accuracy in metastasis prediction.              |
| [17]      | ANN + Deep Learning (IoMT) | High accuracy in predicting breast cancer stages using the IoMT framework. |
| [20]      | ML regression models       | Accurately predicted healthcare costs for breast cancer patients.          |

|      |                               |                                                                        |
|------|-------------------------------|------------------------------------------------------------------------|
| [16] | Deep Learning (CNN, hybrid)   | DL models showed strong accuracy in cancer prognosis.                  |
| [21] | DeepXplainer (DL + XAI)       | High diagnostic accuracy and improved interpretability in lung cancer. |
| [18] | Dimensionality reduction + ML | Autoencoders improved model stability and performance.                 |
| [24] | ML classifiers                | Reported strong accuracy in cancer prediction tasks.                   |
| [6]. | Explainable ML                | Achieved reliable predictions for hospital length of stay.             |

**Table 2.1: Related Works**

### 2.3.1 Support Vector Machine (SVM)

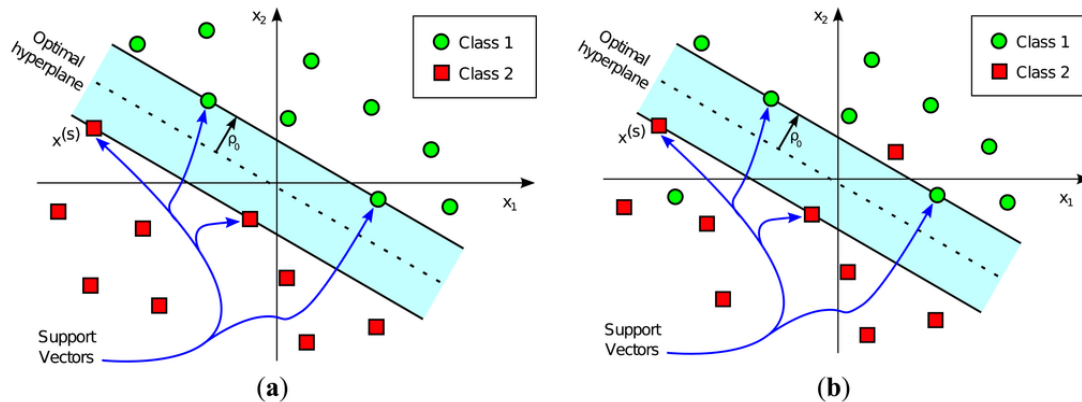


**Figure 2.2: Representation of a Support Vector Machine (SVM) classifier [3]**

SVM, as presented in Figure 2.2 is a supervised classification algorithm which classifies data based on the optimum hyperplane. SVM has also performed superbly in the case of breast cancer prediction, as this tool can work with high-dimensional data. Researchers found that SVM achieves an accuracy of over 90%, hence excelling against the less involved models. It is limited by the cost of computation, utilising huge amounts of data [19] [22].

### 2.3.2 Random Forest (RF)

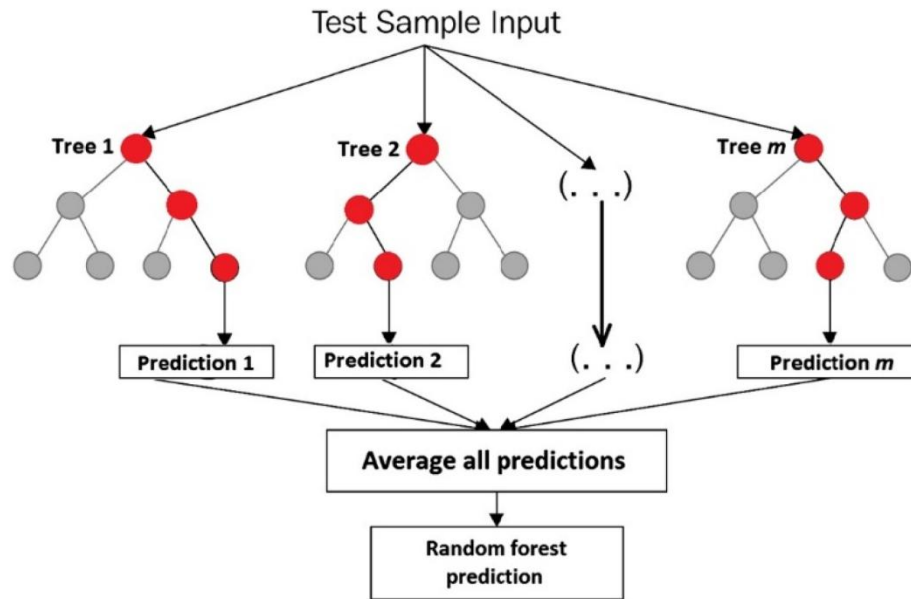


**Figure 2.3: RF prediction algorithm [24]**

Figure 2.3 depicts the Random Forest is an ensemble learning method, which combines many decision trees to enhance accuracy and minimise overfitting. Research works demonstrated that RF was one of the most effective models to classify breast cancer, attaining an accuracy of 92-95% [23] [8]. Its attribute importance option increases the interpretability and therefore is popularly used in medical decision support [25] [8].

### 2.3.3 Logistic Regression (LR)

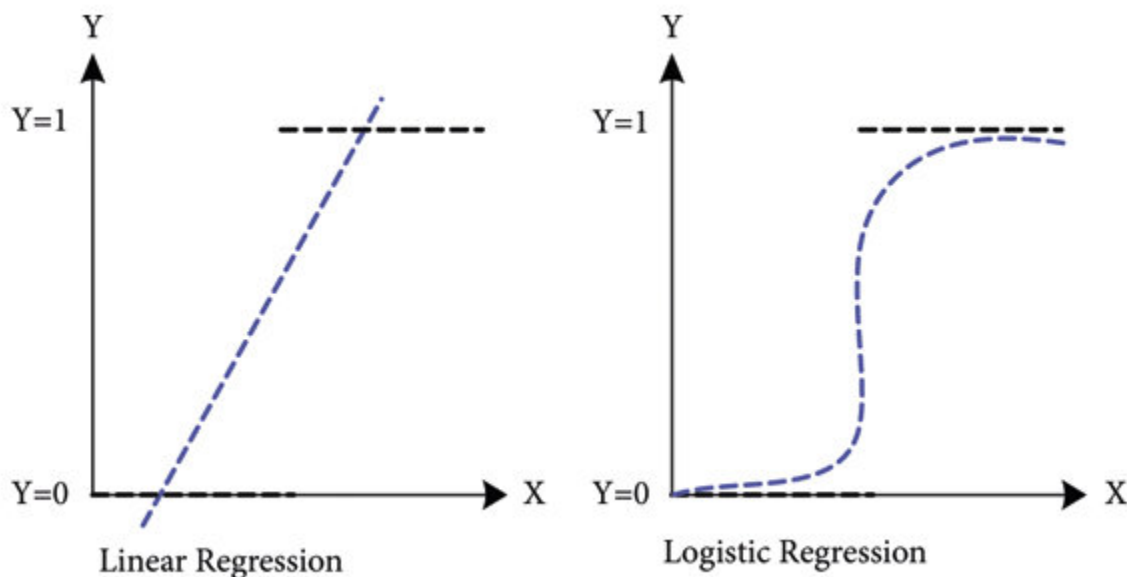


**Figure 2.4: Logistic regression classification model [27]**

Logistic Regression, as presented in Figure 2.4 is the statistical model that is applied to dichotomous classification, which is often applied in breast cancer diagnosis (benign vs. malignant). It works by estimating the probability of an input belonging to a class using a sigmoid function, making it suitable for medical decisions where probability thresholds are important. Despite that it is simplicity, it works well with structured datasets. This was confirmed by researcher, who found that LR has consistent performance and poorer accuracy than RF and SVM [13]. It is also interpretable in nature, but non-linear relationships are problematic [25].

#### 2.3.4 Decision Tree (DT)

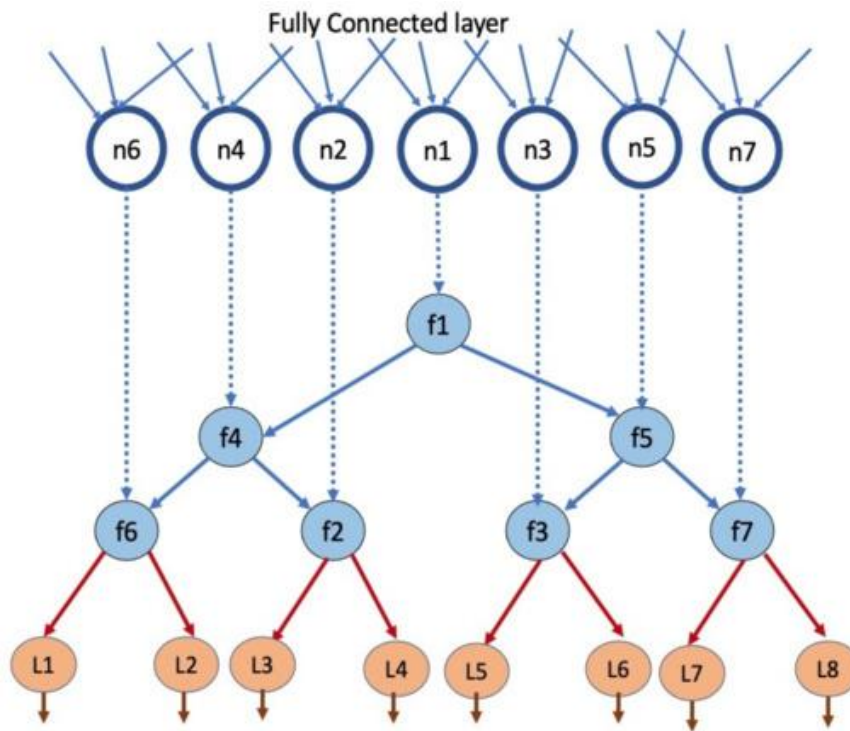


**Figure 2.5: Decision Tree (DT) [7]**

Decision Trees, as presented in Figure 2.5 separate data into branches, dividing attributes, and make understandable rules. They are computationally maybe and are susceptible to overfitting. According to the study, the accuracy of the DTs was moderate over ensemble models, such as RF [26]. Although DTs are applicable in any form of preliminary examination, they are more effective in ensembles to generate some stability and accuracy [27].

### 2.3.5 K-Nearest Neighbour (KNN)

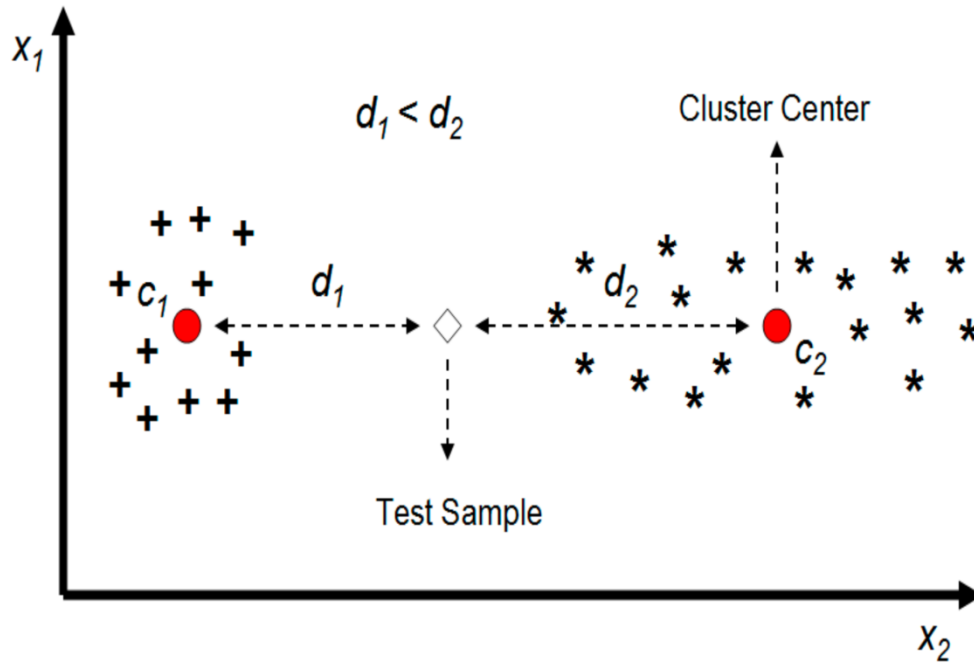


**Figure 2.6: K-Nearest Neighbour (KNN) [40]**

K-Nearest Neighbour (KNN), as presented in Figure 2.6 is a distance classification algorithm which gives labels based on the majority of the closest neighbours. It is easy but efficient in cancer prediction processes. Many authors have discovered that KNN performed fairly well but was somewhat less accurate compared to SVM and RF [30]. Their greatest limitation is that they are computationally inefficient when large volumes of data are involved, and that they are also sensitive to feature scaling [28].

### 2.3.6 Artificial Neural Networks (ANN)

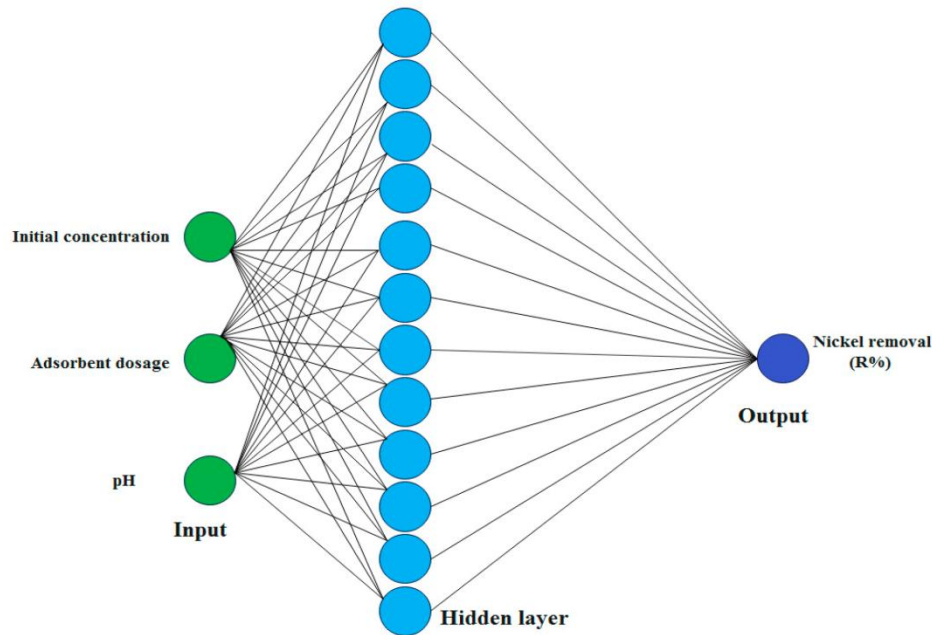


**Figure 2.7: ANN model architecture [29]**

Artificial Neural Networks (ANN), as presented in Figure 2.7 are based on the biological neural systems, which can be used to model complex non-linear relationships. ANN models proved to be very accurate, particularly in detecting the stage and metastasis in breast cancer studies. According to studies, ANN had a performance mark of 93-95% [8] [31]. Although ANN is highly predictive, it is not as interpretable as tree-based models, regardless of the fact that the former is highly predictive [8] [29].

## 2.4 Preprocessing and Feature Engineering in Cancer Prediction

### 2.4.1 Preprocessing Techniques

Preprocessing is a crucial step in preparing biomedical datasets for machine learning. Cancer-related data often contains missing values, noise, or imbalances that may reduce model reliability. Common preprocessing steps include data cleaning, normalisation, and feature scaling. For instance, a study emphasised the role of feature standardisation in improving Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) performance, as both algorithms are sensitive to feature magnitude [32]. Cross-validation and resampling techniques such as k-fold validation and SMOTE (Synthetic Minority Over-sampling Technique) are also applied to minimise bias and address class imbalance in breast cancer datasets. Similarly, [18] highlighted the significance of

balanced training data for ensemble models like Random Forest, ensuring stable classification outcomes.

#### **2.4.2 Feature Selection and Representation**

Feature engineering allows models to identify the most relevant predictors in cancer datasets. The Breast Cancer Wisconsin dataset, for example, contains attributes such as radius, texture, smoothness, concavity, and compactness. The study demonstrated the impact of dimensionality reduction techniques, such as Principal Component Analysis (PCA) and autoencoders, in improving classification efficiency while reducing redundancy [22]. Random Forest models provide inherent feature importance scores, which were shown by [8] to aid interpretability in metastasis prediction. Moreover, deep learning models are capable of automated feature extraction from complex datasets, such as medical images, enhancing predictive performance without manual selection. Effective feature representation thus strengthens accuracy and reduces overfitting, ensuring models generalise better to unseen cases.

#### **2.5 Deep Learning and Hybrid Approaches**

However, deep learning (DL) has been gaining more and more recognition as a cancer prediction method because of its ability to learn highly complex feature representations in large and high-dimensional data automatically. Convolutional Neural Networks (CNNs) are particularly useful with imaging-based potentially cancerous diagnoses, including mammograms, histopathological slides, and radiological scans, in which subtle spatial patterns can hint at malignancy. Connectivity's Clinical and genomic data have also been trained using fully connected deep neural networks and recurrent structured architectures with improved performance on tasks such as progression risk estimation and survival prediction compared to classical classifiers [31].

Here, a key shortcoming of deep learning models is their understanding. Because a clinical choice requires transparency, multiple studies have proposed explainable AI (XAI) models to speak in addition to DL forecasts. SHAP values and saliency maps are among the methods, and the recently introduced DeepXplainer enables clinicians to introspect attribute contributions/ontological aspects to a given prediction. The increasing acknowledgement of such integration of interpretability tools as being a prerequisite of winning physician trust and regulatory acceptance is becoming established [34].

Deep-learning feature extraction that is complemented by classical machine learning algorithms (like hybrid approaches) is promising as well. An example is the application of CNNs to extract

imaging features and employing ensemble reclassification models: XGBoost or random forests, obscuring any difference between individual models or combined models. Likewise, the dimensionality reduction by using the autoencoders has been integrated with standard classifiers to enhance efficiency and to alleviate over-fitting of smaller data sets. Such hybrid pipelines provide a trade-off between accuracy and interpretability, as well as scalability with a variety of healthcare data [32].

## **2.6 Applications Beyond Diagnosis**

### **2.6.1 Prognosis and Treatment Planning**

While early diagnosis remains a primary focus, machine learning (ML) applications in cancer care extend far beyond detection. Predictive models for disease progression and survival allow clinicians to stratify patients into risk groups, tailoring follow-up schedules and therapeutic interventions accordingly [40]. Similarly, ML frameworks assist in metastasis detection and staging, providing decision support for treatment planning and long-term care management [8]. Risk prediction algorithms also forecast surgical complications and medication side effects by analysing preoperative factors and patient histories, enabling preventive measures that enhance patient safety [33].

### **2.6.2 Healthcare Economics and Resource Management**

ML is increasingly applied in healthcare economics, such as predicting the cost of breast cancer care, which helps health systems plan budgets and allocate resources efficiently, ultimately reducing financial burdens for patients and providers [34]. Explainable ML models can also predict hospital length-of-stay, aiding in bed allocation and alleviating overcrowding [6]. These applications demonstrate that ML in cancer research transcends simple classification of malignancy. By supporting prognosis, economic forecasting, complication risk assessment, and operational efficiency, ML contributes to the full spectrum of cancer management—from diagnosis to sustained patient care [39].

## **2.7 Federated Learning and Data Privacy in Cancer Prediction**

The exchange of information across repositories is also among the complications which just do not fade in the hardcore attempts to derive plausible patterns to forecast cancer. There is a tendency to store cancer information in closed facilities in hospitals due to patient confidentiality, a deficiency of information regarding ethics, and legal restrictions on medical archives. This fragmentation is likely to limit training data diversity and representativeness and lead to models

that will not generalise across or between populations. Federated learning (FL) is a more promising paradigm, in order to handle this challenge. Instead of centralising sensitive data, FL can be applied to a group of several healthcare institutes where multiple markers can be trained and maintain localised data jointly. Sharing done in Elizabetha models shares only in model parameters, so privacy sharing and probability of data leakage are minimised [35].

The concept of federated approaches has shown promise in cancer predictions with tasks because, unlike larger models, they can be developed in large volumes without security concerns. One such integration practice is the approach of prognosis and diagnostic modelling of diseases that FL can integrate heterogeneous datasets of clinical sites and thus enhancing model robustness and external validity. The latter approach will particularly work well with rare subtypes of cancers where limited datasets on any single validation site cannot be modified to useful models [23].

In this case, there are some technical and organisational problems. FL models can be inefficient and less accurate because of communication overhead between involved nodes, the distribution heterogeneity of data and disadvantages in preprocessing pipelines. Furthermore, equality and non-bias within facilities are a burning issue since the already trained models, working only with data that depend on high-resource hospital networks, are not necessarily equally bedroom in the cases that are underrepresented. The other critical issue raised by scientists is explainability by the FL framework that should be guaranteed about clinical trustworthiness [2].

Lastly, it is possible to mention federated learning, which is going to become a significant innovation in the privacy-protected pathway of cancer prediction. FL introduces an incentive that is more generalisable and equitable that enabling the utilisation of collaborative training across more than two centres. Nonetheless, to be effectively implemented, certain basic challenges must be addressed to address technical challenges, harmonise data controls, and openness of decision-making processes [37].

## **2.8 Literature Gap**

Despite the many reports that have been making use of machine learning to predict cancer, there are notable gaps that exist. Lots of publications acknowledge high accuracies with models like Support Vector Machines, random forests, and neural networks, but their generalisability is often doubted, because many models depend on small datasets, like the Breast Cancer Wisconsin dataset. Although these datasets are helpful as benchmark data, they are not diverse in demographic

and clinical characteristics, and therefore, the results cannot be used to generalised the results to larger populations.

The other bleeding weakness is related to the imbalance in datasets. Malignant samples are normally underrepresented, and the results of these models are generally biased to predict benign cases. Not many investigations have been conducted on this issue systematically with a method of high resampling or through cost-sensitive learning. In addition, accuracy is often cited as the primary evaluation measure, whereas less focus is on clinically significant ones, like recall or F1-score, that directly influence the possibility of minimising false negatives.

Another issue that is not compromised is interpreting it. Deep learning models have a great predictive capability but serve as black boxes and thus are not trusted by clinicians. Explainable AI systems have been developed, but these have not been increasingly utilised in conjunction with breast cancer prediction. Lastly, a variety of comparative studies end before making any action-related recommendations on the trade-offs between interpretability versus computational cost, versus predictive accuracy. Hence, it is in the light of literature that there is a strong desire to perform more extensive, comparative assessments of machine learning models on heterogeneous datasets, balanced measures, and place more stress on interpretability to enable clinical adoption.

## Chapter 3: Methodology

### 3.1 Introduction

This chapter describes the methodology used to accomplish the objective of forming a comparative analysis of machine learning models to give a reliable breast cancer prediction. It gives a description of the philosophical position, research design, research methods and research strategy used in the study. The chapter also describes the tools, techniques, datasets and machine learning models to be experimented with. Also, it describes the preprocessing of the dataset that will be used, the reasons behind choosing the algorithms and libraries, and ethical concerns on which the work is based. Collectively, these three methodological options guarantee rigour, replicability, and validity in conducting the research.

### 3.2 Research Philosophy, Design and Approach

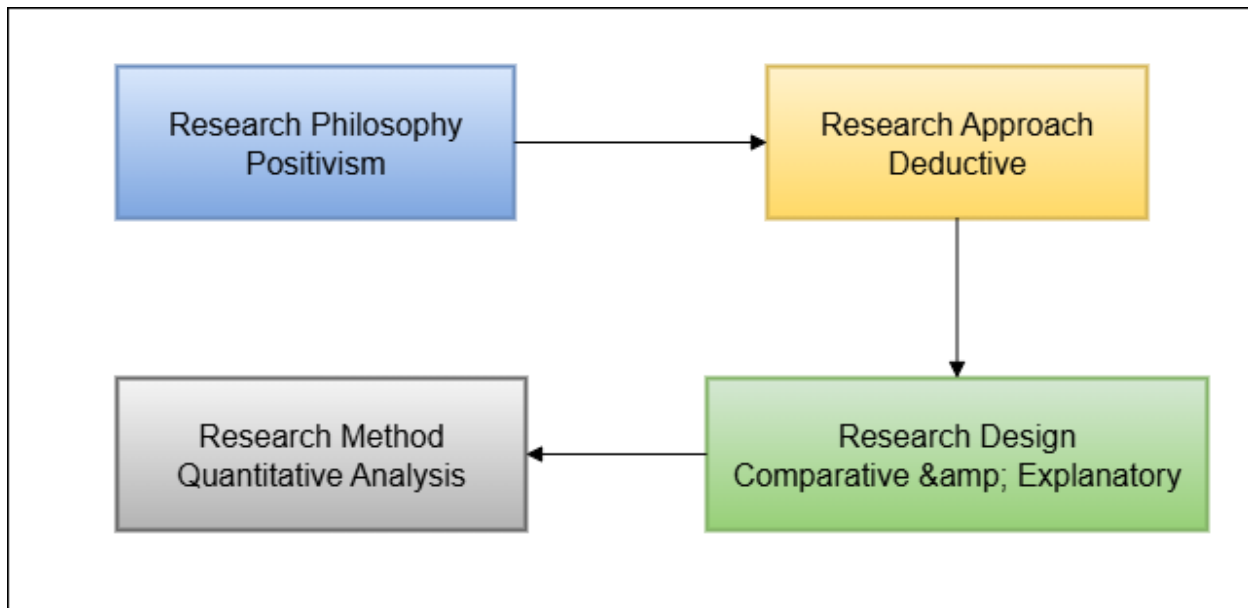


**Figure 3.1: Research Philosophy, Design and Approach**

(Source: Self-Created)

The study is based on a positivist philosophy, since the research follows objective, measurable evidence, which was achieved through machine learning experimentation. A deductive approach was used, where developed theories and models on the prediction of cancers were used, and their assumption tested with real data to substantiate the validity. It designs its research as a comparative and explanatory study and was meant to explore the performance difference between different models. Also, a quantitative orientation was chosen as the analysis is based on numerical

parameters like accuracy, precision, recall, and F1-score, which enable any model to be systematically evaluated and compared.

### 3.3 Tools and Techniques

The approach is a combination of computational tools and structured approaches in line with the research objectives. The most fundamental techniques are summarised in Table 3.1.

| Technique              | Purpose                                                                                |
|------------------------|----------------------------------------------------------------------------------------|
| Data Preprocessing     | Cleaning, resizing, and normalising ultrasound images for consistency.                 |
| Feature Extraction     | Capturing patterns and attributes from images to support classification [8].           |
| Data Splitting         | Dividing the dataset into training, validation, and test sets for fair evaluation.     |
| Model Training         | Applying selected machine learning algorithms to training data.                        |
| Performance Evaluation | Assessing models using metrics such as accuracy, precision, recall, and F1-score [41]. |

**Table 3.1: Applied Techniques in Methodology**

#### 3.3.1 Data Preprocessing

Preprocessing makes the data clean, and modelling is now good. Breast ultrasound imaging in this study was normalised and resized to generate uniform samples among the samples. The step eliminated the scaling inconsistencies and also improved the performance of feature scale-sensitive algorithms.

#### 3.3.2 Feature Extraction

The processed ultrasound images were used to extract relevant features of possible tumours: shapes, textures and boundaries. These characteristics comprise the input vectors, which allow the models to distinguish between benign, malignant and normal cases [9].

#### 3.3.3 Data Splitting

The data was split into learning, validation and testing. This randomisation made sure that evaluations were done on hidden content, thus minimising bias and overfitting. K-fold cross-validation has been used to enhance the strength of the performance estimates [43].

### 3.3.4 Model Training

Optimised hyperparameters were used to train machine learning models, namely Support Vector Machine (SVM), Random Forest (RF), Logistic regression, Decision tree, K-nearest neighbour (KNN) and Artificial Neural Networks (ANN). The models were also able to train in order to know the decision boundaries where they could classify correctly.

### 3.3.5 Performance Evaluation

Assessment was done by calculating accuracy, precision, recall, F1-score. With the clinical significance of minimising false negatives, this set of metrics offered a more multi-dimensional perspective of model effectiveness, specifically focusing on sensitivity and specificity [42]. In this case, to ensure that the appraisals of the performance of the models are sound, several evaluation measures will be employed to measure the performance of the models:

- Accuracy: Percentage accuracy of classification of cases of abuse.
- Accuracy: Percentage of successes among what have been foretold to be positive.
- Recall (Sensitivity): True positives that are perceived.
- F1: Precision and Recall.
- Precision–Recall (PR) Curve: This metric evaluates the balance between precision and recall, focusing on the model’s ability to detect positive cases. It is especially useful for imbalanced datasets where positive class performance is more critical.

By the multiple existence of metrics, the analysis avoids over-reliance on accuracy, likely to be unreliable on skewed data.

| Step No. | Process                    | Description                                                                                          |
|----------|----------------------------|------------------------------------------------------------------------------------------------------|
| 1        | Define Research Philosophy | Adopted positivism to ensure objective, measurable, and replicable findings.                         |
| 2        | Select Research Approach   | A deductive approach was chosen to test existing machine learning theories against dataset evidence. |
| 3        | Determine Research Design  | A comparative, explanatory, and quantitative design was selected to evaluate multiple models [40].   |
| 4        | Data Collection            | Used the secondary dataset (Breast Cancer ultrasonic dataset) for experimentation.                   |

|    |                             |                                                                                                          |
|----|-----------------------------|----------------------------------------------------------------------------------------------------------|
| 5  | Data Preprocessing          | Handled missing values, normalised/standardised features, and applied feature selection.                 |
| 6  | Model Selection             | Choose six algorithms: Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and ANN (Sun, 2022). |
| 7  | Model Implementation        | Implemented models using Python libraries (Scikit-learn, TensorFlow). Hyperparameters tuned.             |
| 8  | Model Training & Validation | Applied train-test split and k-fold cross-validation to ensure reliability [8].                          |
| 9  | Performance Evaluation      | Compared models using Accuracy, Precision, Recall, F1-score, and Precision–Recall (PR) Curve.            |
| 10 | Comparative Analysis        | Analysed results side by side to determine the most reliable algorithm.                                  |
| 11 | Draw Conclusions            | Interpreted findings to recommend the best model and reflect on limitations.                             |

**Table 3.2: Method Steps**

### 3.4 Data Collection

The dataset in this study was the Breast Ultrasound images Dataset downloaded on Kaggle (Source: <https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset>). It entails the labelled ultrasound imaging in the normal, benign, and malignant categories. This dataset was selected because it is pertinent to clinical practice and is mostly applicable in predictive modelling research. Besides a primary dataset, the secondary data sources were used as additional sources of literature review and theoretical underpinning. These were peer-reviewed journals in databases like Google Scholar, MDPI, Elsevier, and IEEE Xplore. These sources, in combination, present both theoretical and empirical bases to the research.

### 3.5 Used Libraries and Machine Learning Models

| Library       | Purpose / Why Used                                                       |
|---------------|--------------------------------------------------------------------------|
| <b>NumPy</b>  | For efficient numerical operations and handling multidimensional arrays. |
| <b>Pandas</b> | For structured data manipulation, cleaning, and tabular processing.      |

|                           |                                                                                     |
|---------------------------|-------------------------------------------------------------------------------------|
| <b>Matplotlib</b>         | For plotting graphs and visualising training and evaluation results.                |
| <b>Seaborn</b>            | For statistical visualisation, including heatmaps of confusion matrices.            |
| <b>Scikit-learn</b>       | For implementing classical ML models (SVM, RF, LR, DT, KNN) and evaluation metrics. |
| <b>TensorFlow / Keras</b> | For building and training Artificial Neural Networks (ANN) effectively.             |
| <b>OS / Glob</b>          | For file and directory handling to load breast ultrasound images.                   |
| <b>OpenCV</b>             | For image preprocessing (resizing, grayscale, normalisation).                       |

**Table 3.3: Used Libraries**

The methodology had been worked out with the help of several Python libraries. The use of NumPy and Pandas helped to manage and perform math on data. Data visualisation and performance plotting were done using Matplotlib and Seaborn. The scikit-learn supports implementations of classical machine learning models like Logistic Regression, Decision trees, Random Forest, SVM or KNN as well as preprocessing and model evaluation. In deep learning, TensorFlow and Keras were used to construct and train Artificial Neural Networks effectively.

| <b>Model</b>                        | <b>Why Selected / Key Strength</b>                                                  |
|-------------------------------------|-------------------------------------------------------------------------------------|
| <b>Logistic Regression (LR)</b>     | Provides a simple and interpretable baseline for binary classification.             |
| <b>Decision Tree (DT)</b>           | Transparent model, useful for initial exploration of feature importance.            |
| <b>Random Forest (RF)</b>           | A robust ensemble method reduces overfitting, high accuracy in biomedical datasets. |
| <b>Support Vector Machine (SVM)</b> | Effective in high-dimensional spaces, strong accuracy for cancer classification.    |

|                                        |                                                                         |
|----------------------------------------|-------------------------------------------------------------------------|
| <b>K-Nearest Neighbour (KNN)</b>       | Simple distance-based classifier, useful for comparison.                |
| <b>Artificial Neural Network (ANN)</b> | Captures complex non-linear patterns, strong performance on image data. |

**Table 3.4: Developed Models**

The reason behind the selection of machine learning models was the soft power of these models. Baseline interpretability was provided by the use of Logistic Regression and Decision Trees. Both Random Forest and SVM were very strong and highly accurate with structured biomedical data. KNN would be a comparative classifier used as a distance-based one [33]. Lastly, ANN was added because it has the capacity to memorise complicated nonlinear patterns of medical pictures. A comparison between these models worked to guarantee a just comparative test, which was used to determine the most trusted algorithm in breast cancer forecasting. The choice of different models gave us ideas about trade-offs, interpretability, accuracy, as well as computational complexity.

### 3.6 Theoretical Framework

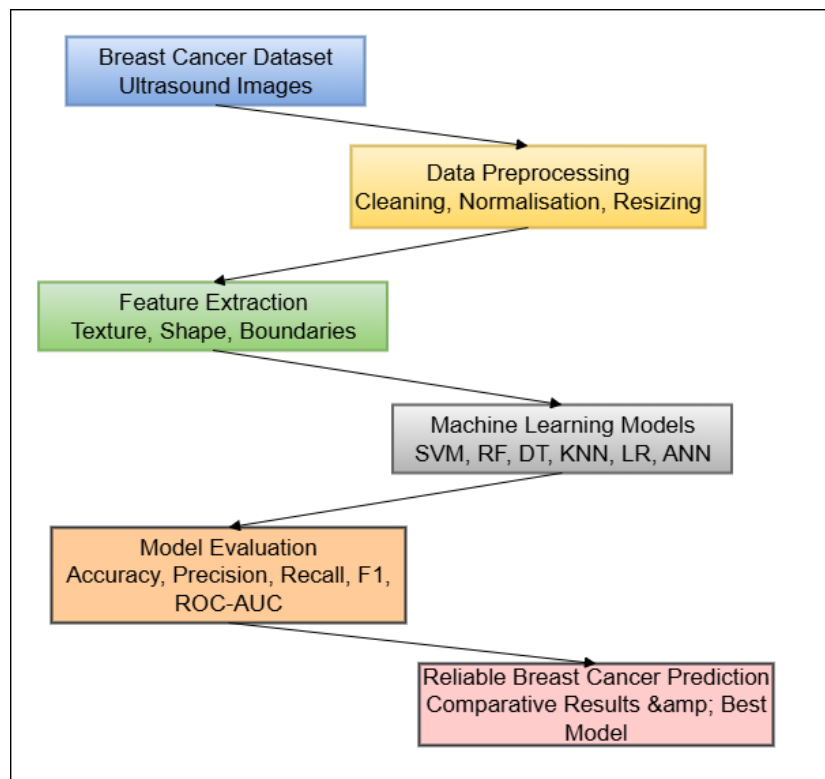


**Figure 3.2: Theoretical Framework**

(Source: Self-Created)

The theoretical framework adopted by this study will be based on the use of supervised machine learning in medical prediction. Supervised learning allows the models to learn decision limits based on the labelled data sets of benign and malignant tumour samples and make predictions on the unknown data. The framework integrates preprocessing and feature extraction theories, which are used to ensure quality of the inputs, ensemble and kernel-based theory learning, which can guarantee robustness and evaluation frameworks driven by measures such as accuracy, recall and F1-score. Combined, the theories help in a comparative analysis of several algorithms, which brings the student onto the same page in determining stronger models to use in predicting breast cancer.

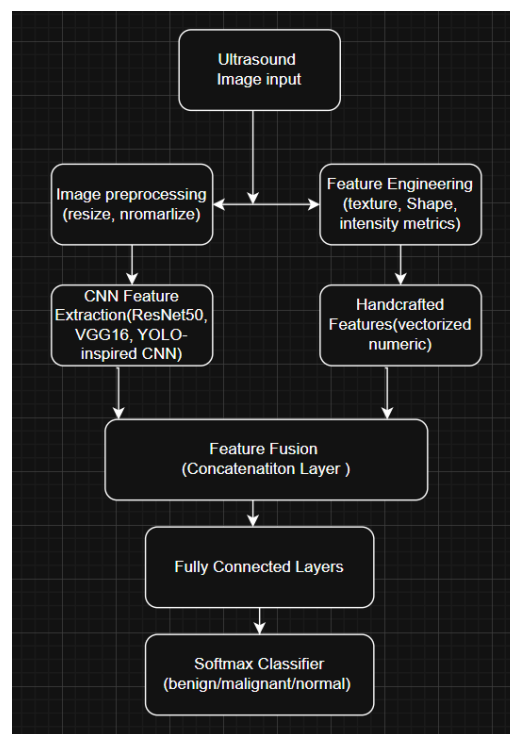


**Figure 3.3: Theoretical Framework for hybrid CNN Model**

(Source: Self-Created)

The proposed Hybrid framework integrates deep learning and handcrafted feature extraction for ultrasound-based breast cancer prediction. The process begins with image preprocessing to enhance quality, followed by dual feature extraction CNN based deep features from YOLO, VGG16, ResNet50 and statistical handcrafted features (texture, shape, intensity). These are combined through feature fusion, producing a through fully connected layers and a SoftMax classifier to categorise cases as benign, malignant, or normal. This hybrid approach leverages both spatial and statistical information, improving diagnostic accuracy and model interpretability.

### 3.7 Project Plan

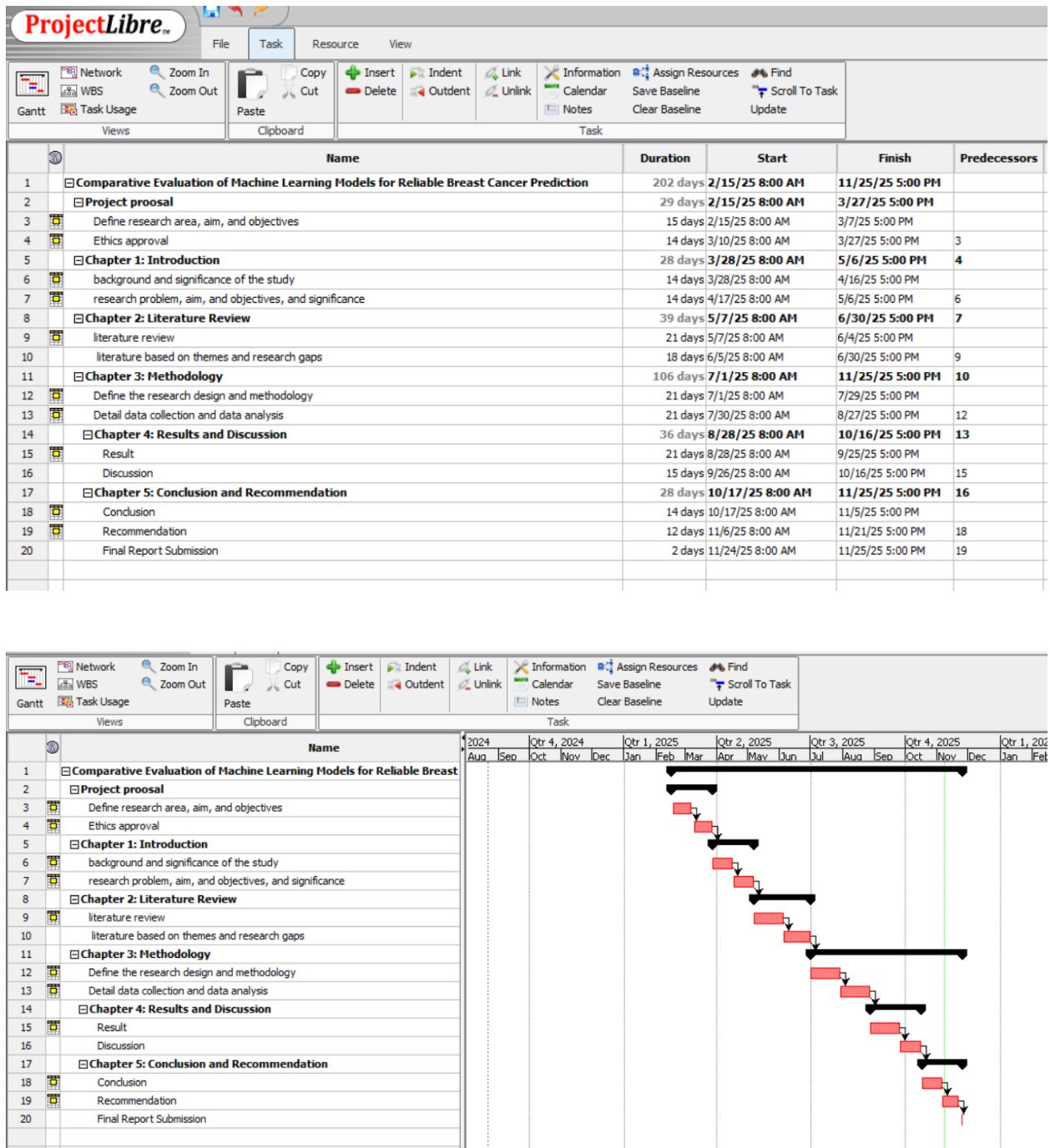


Figure 3.4: Project Plan  
(Source: Self-Created)

The provided project plan represents a well-organized chronological timeline of a research project that will be named Comparative Evaluation of Machine Learning Models to Reliable Breast Cancer Prediction. It has been arranged in Work Breakdown Structure format, and the activities are arranged between February 2025 and November 2025. The plan will start with the phase of proposal of the project in which the ethics approval, aim, objectives and area of research will be established. This occurs within a period of about a month. The next step is Chapter 1 (Introduction) which dwells on background, significance, and formation of research problem.

The following plan outlines Chapter 2 (Literature Review) as the second chapter that will follow the introductory chapter; however, Chapter 2 has worked out two main tasks to be performed: literature review, in general, and theme-and-gap identification, in particular. Chapter 3 (Methodology) takes the longest time as there is complexity therein. It incorporates specification of the methodology, which is followed by elaborate data collection and analysis processes, which is in line with a viable ML experiment workflow. Chapter 4 (Results and Discussion) is generated in the later phases of the project with results generation and interpretation being separate time blocks. Chapter 5 (Conclusion and Recommendation) is the last academic chapter that summarises the outcomes and cogitates recommendations.

The last step of the project plan is the submission of final report. The total number of days covered is 202, which begins on 15 February 2025 and ends on 25 November 2025. The structure only represents explicit predecessor dependencies, which illustrate a logical dependency between the tasks and exhibits a step-by-step, research-based approach, starting with preparation to the final reporting.

### **3.8 Ethical Consideration**

Ethical standards in the use of secondary data were followed in the research to the letter. The data resource was publicly accessible on Kaggle, and there were no issues with patient privacy and consent. The dataset did not contain any personal identifiers [21]. The presence of academic integrity was ensured correctly by citing all data sources and references. Additionally, cross-validation, along with various measures of performance, served to mitigate this bias during model evaluation so that there were no risks of being misrepresented. Ethical responsibility played a part

in the reporting of the results, and the findings are presented transparently. There is no exaggeration of claims about clinical applications.

## Chapter 4: Results Evaluation and Discussion

### 4.1 Introduction to Results

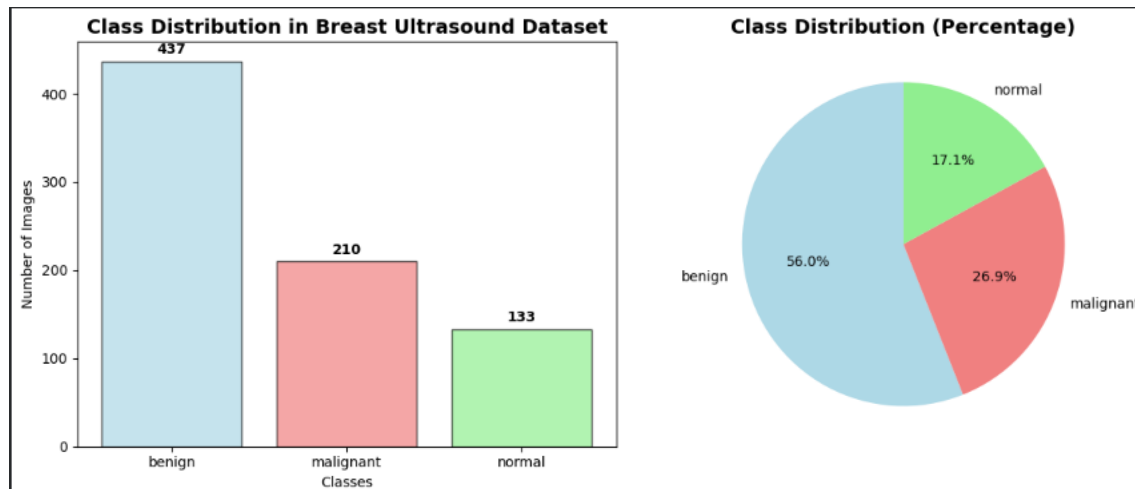


**Figure 4.1: Class Distribution**

(Source: Developed Using Python)

This paper compares machine learning systems to projects on breast cancer prediction based on the dataset of BUSI ultrasound (benign 56, malignant 27, and normal 17), with 780 images in total. Four architectures were evaluated, namely a custom YOLO-based CNN, VGG16 transfer learning, and ResNet50 transfer learning, and the Hybrid CNN + Handcrafted Features model. The analysis involves the preprocessing, performance metrics, feature extraction, model training, comparative evaluation, and clinical relevance. There was a fluctuation in the imbalance of datasets, but it represents the normal distribution of clinical conditions in the real world, where benign conditions rank highest, malignant conditions are fewer, and normal conditions are scarce. See what conclusions offer information on the efficiency of models, errors corrected when establishing the diagnosis, and the way it can be applied in medical imaging.

### 4.2 Results and Analysis

#### 4.2.1 Data Pre-processing and Augmentation Results

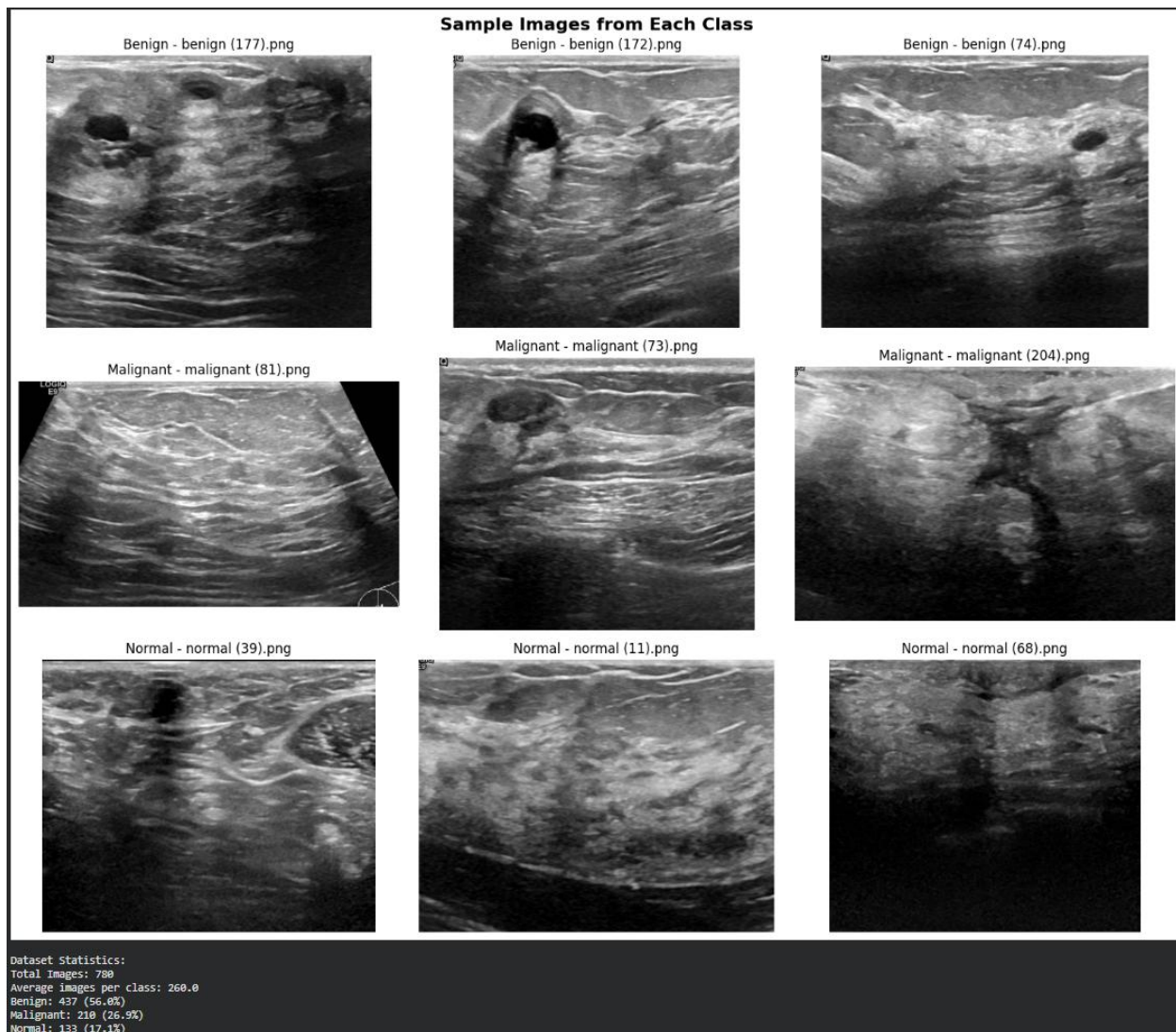


**Figure 4.2: Sample Images**

(Source: Developed Using Python)

The data preprocessing pipeline used advanced procedures in order to improve the image quality and to improve the generalisation of models. Image contrast and detail were enhanced with Contrast Limited Adaptive Equalisation (CLAHE) 2.0 clip-limit and an  $8 \times 8$  grid, which resolved the presence of noise in the image and improved the ultrasound image quality and clarity of images [32]. Available on approximately 60% of images, one mask-based region of interest (ROI) extraction localised lesions by switching masks and cropping interagency of interests localised models and focused their models on one diagnostically salient area.

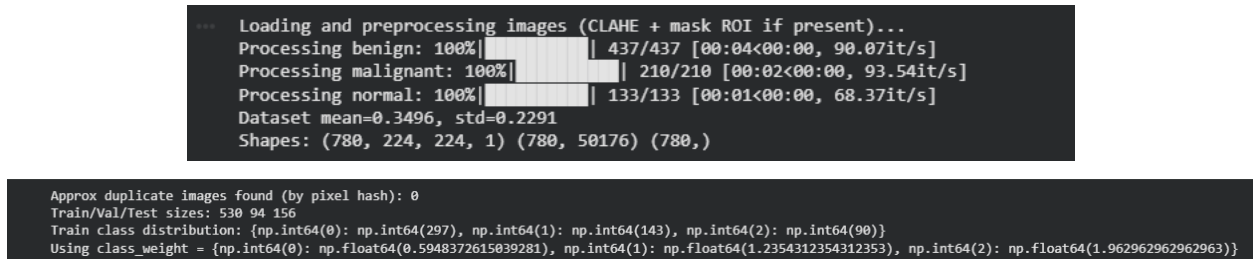


**Figure 4.3: DataFrame after Pre-processing**

(Source: Developed Using Python)

The dataset-specific mean (0.3496) and standard deviation (0.2291) were used as standardisation and guaranteed constant training and constant input distributions, and median blur filtering (3x3 kernel) eliminated noise without damaging edges that were important in diagnosis. To obtain additional robustness of the models, it was also allowed that controlled variability in the form of rotations ( $\pm 20^\circ$ ), width/height shift (10%), flips, zoom (recognition  $\pm 15\%$ ) and brightness variation (0.8 - 1.2) was introduced. These add-ons were realistic clinical variation imaging, which enhanced model generalisation without distorting the diagnostic characteristics. The preprocessing pipeline, in general, took into account the challenges of contrast, imbalanced classes, limited sample, and noise, and it finally aided in proper and clinically useful breast cancer detection based on ultrasound images.

#### 4.2.2 Training Dynamics and Convergence Analysis

The Four models presented unique convergence behaviours during the training process. The YOLO-like CNN was trained with stable convergence, taking a maximum of 25 epochs with the assistance of early stopping to avoid overfitting. It lost its curves in a smooth mode, and this indicated a good selection of learning rate and batch normalisation. VGG16 and ResNet50 transfer learning models, conversely, required fewer epochs to converge, embracing the ultrasound domain in 10 epochs, and may be used in converting to the ultrasound domain as these feature extractors are trained [5]. Fine-tuning needed to be planned carefully, not to experience catastrophic forgetting, but to have the opportunity to adapt to the tasks. Weighting (benign 0.59, malignant 1.24, normal 1.96) solved the issue of class imbalance and provided equal opportunities on the optimisation step. Adaoptimiser using the seed rate of 0.001 was supportive to every model, whereas ReduceLROnPlateau (factor 0.5, patience 4) allowed fine-tuned adjustments in the further stages. This adaptive outcome was especially valuable to transfer learning models that trade off maintained characteristics with medical specialised learning. Stratified sampling was used to

divide the data (80%-15%-15) (train-validation-test) to ensure that the proportion of classes is not disproportionate and that there is no bias [5]. Stopping patience of 8 epochs and best weight restoration was an additional measure against overfitting that still allowed the best generalisation conditions, without causing overfitting.

## 4.3 Model Evaluation

### 4.3.1 Performance Metrics Overview

On a test set of 156 ultrasound images (87 benign, 42 malignant, 27 normal), four models—YOLO-inspired CNN, VGG16, ResNet50, and the proposed Hybrid CNN + Handcrafted Features—were evaluated based on accuracy, precision, recall, F1-score, and Precision–Recall (PR) analysis. None of the results indicated any significant differentiation, although the Hybrid VGG16+ Handcrafted model achieved the highest accuracy of 88% hybrid ResNet50 (87%) and hybrid YOLO (85%). The standalone YOLO-inspired CNN achieved 85.90% VGG16 (81.41%) and ResNet50 (78.21%). The hybrid model achieved performance comparable to the transfer learning model, demonstrating that combining handcrafted statistical features with deep CNN-based features can enhance discriminative representation while maintaining training stability. Overall, these results confirm that deep feature extraction remains the most effective for ultrasound classification, while hybrid fusion approaches offer a balanced trade-off between accuracy and interpretability.

### 4.3.2 Detailed Analysis of Model Performance

- YOLO-Inspired CNN Performance

```

=== YOLO-inspired CNN Summary ===
Accuracy: 0.8590
Macro Precision: 0.8717 | Macro Recall: 0.8664 | Macro F1: 0.8689
Weighted Precision: 0.8575 | Weighted Recall: 0.8590 | Weighted F1: 0.8581

Classification Report:

```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Benign       | 0.87      | 0.89   | 0.88     | 87      |
| Malignant    | 0.75      | 0.71   | 0.73     | 42      |
| Normal       | 1.00      | 1.00   | 1.00     | 27      |
| accuracy     |           |        | 0.86     | 156     |
| macro avg    | 0.87      | 0.87   | 0.87     | 156     |
| weighted avg | 0.86      | 0.86   | 0.86     | 156     |

**Figure 4.4: YOLO-Inspired CNN Model Performance**

(Source: Developed Using Python)

CNN, inspired by YOLO, had found good results with a macro F1-score of 0.8689, precision of 0.8717, and recall of 0.8664, indicating even performance when considering the three classes. With benign cases, the recall was 0.89 and the precision was 0.87, which is found to be effective in

detecting a huge portion of benign tumours. The challenges concerning clinical difficulties were reflected in a lower malignant detection, 0.71 recall, and 0.75 precision. It had good generalisation because of its architecture, which employed the use of layered convolution, batch normalisation, dropout and global average pooling.

- **ResNet50 Transfer Learning Model**

```
=== ResNet50 Transfer Learning Summary ===  
Accuracy: 0.7821  
Macro Precision: 0.8081 | Macro Recall: 0.8451 | Macro F1: 0.8105  
Weighted Precision: 0.8291 | Weighted Recall: 0.7821 | Weighted F1: 0.7873  
  
Classification Report:  
      precision    recall  f1-score   support  
  
   Benign      0.92      0.68      0.78        87  
  Malignant    0.57      0.86      0.69        42  
    Normal    0.93      1.00      0.96        27  
  
   accuracy          0.78        156  
  macro avg      0.81      0.85      0.81        156  
 weighted avg    0.83      0.78      0.79        156
```

**Figure 4.5: ResNet50 Transfer Learning Model Performance**

(Source: Developed Using Python)

The ResNet50 language model with transfer learning produced good results with a macro F1-score of 0.8105, Precision of 0.8081 and a macro recall of 0.8451, respectively, which are balanced. It had skip connections and residual blocks, which further added features to the networks by alleviating vanished gradients in deep networks. It did well with benign cases (0.68 recall, 0.92 precision); however, with malignant cases, it showed moderate sensitivity (0.86 recall, 0.57 precision). In the usual cases, it obtained an ideal recall (1.00), and the precision is 0.93. ResNet50 was the second most multicast-proven powerful third.

- **VGG16 Transfer Learning Model**

```

=== VGG16 Transfer Learning Summary ===
Accuracy: 0.8141
Macro Precision: 0.8220 | Macro Recall: 0.8560 | Macro F1: 0.8318
Weighted Precision: 0.8388 | Weighted Recall: 0.8141 | Weighted F1: 0.8186

Classification Report:

```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Benign       | 0.92      | 0.76   | 0.83     | 87      |
| Malignant    | 0.62      | 0.81   | 0.70     | 42      |
| Normal       | 0.93      | 1.00   | 0.96     | 27      |
| accuracy     |           |        | 0.81     | 156     |
| macro avg    | 0.82      | 0.86   | 0.83     | 156     |
| weighted avg | 0.84      | 0.81   | 0.82     | 156     |

**Figure 4.6: VGG16 Transfer Learning Model Performance**

(Source: Developed Using Python)

The VGG16 transfer learning model exhibited a different picture with the second highest accuracy (81.41) and maximum macro recall (0.8560) macro F1-score of 0.831805, Precision of 0.8220, and indicates high performance in sensitivity between classes. It had a recall of 0.76 and a precision of 0.92 in benign cases and a maximum recall of 0.81 with a precision of 0.62 Normal Precision **0.93**, Recall **1.00** in cases where it is malignant, meaning that it is sensitive rather than precise. It had 16 sequential layers of the 3x3 filter, which made this effective in extracting features, but the absence of residual connections restricted higher-order relationships.

- Hybrid (YOLO + Handcrafted):

```

=== Hybrid (YOLO + Handcrafted) Summary ===
Accuracy: 0.8462
Macro Precision: 0.8588 | Macro Recall: 0.8588 | Macro F1: 0.8588
Weighted Precision: 0.8462 | Weighted Recall: 0.8462 | Weighted F1: 0.8462

Classification Report:

```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Benign       | 0.86      | 0.86   | 0.86     | 87      |
| Malignant    | 0.71      | 0.71   | 0.71     | 42      |
| Normal       | 1.00      | 1.00   | 1.00     | 27      |
| accuracy     |           |        | 0.85     | 156     |
| macro avg    | 0.86      | 0.86   | 0.86     | 156     |
| weighted avg | 0.85      | 0.85   | 0.85     | 156     |

**Figure 4.7: Hybrid (YOLO + Handcrafted) Model Performance**

(Source: Developed Using Python)

The Hybrid CNN model, integrating YOLO- extracted deep features with handcrafted statistical descriptions, achieved an accuracy of 84.62%, with macro precision of 0.8588, macro recall of

0.8588, and macro F1-score of 0.8588. These results demonstrate that the hybrid fusion effectively balanced sensitivity and specificity across all classes. For benign cases the model obtained a precision of 0.86 and recall of 0.86, showing strong recognition of non-malignant tumours. The malignant class achieved a recall of 0.71 and precision of 0.71, reflecting improved sensitivity than standalone CNN models. The normal class achieved near-perfect result recall of 1.00 and precision of 1.00, confirming reliable detection of healthy tissues. Overall, the hybrid approach demonstrated stable and generalised performance, effectively combining deep and handcrafted features for improved breast cancer detection,

- Hybrid (VGG16+ Handcrafted):

```

=== Hybrid (VGG16 + Handcrafted) Summary ===
Accuracy: 0.8782
Macro Precision: 0.8909 | Macro Recall: 0.8820 | Macro F1: 0.8860
Weighted Precision: 0.8766 | Weighted Recall: 0.8782 | Weighted F1: 0.8770

Classification Report:

```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Benign       | 0.88      | 0.91   | 0.89     | 87      |
| Malignant    | 0.79      | 0.74   | 0.77     | 42      |
| Normal       | 1.00      | 1.00   | 1.00     | 27      |
| accuracy     |           |        | 0.88     | 156     |
| macro avg    | 0.89      | 0.88   | 0.89     | 156     |
| weighted avg | 0.88      | 0.88   | 0.88     | 156     |

**Figure 4.8: Hybrid (VGG16 + Handcrafted) Model Performance**

(Source: Developed Using Python)

The Hybrid VGG16 model achieved an accuracy of 87.82% with macro precision of 0.8909, macro recall of 0.8820, and macro F1 of 0.8860. For benign cases, precision of 0.88 and recall of 0.91, while malignant cases achieved 0.79 precision and 0.74 recall. The normal class performed best with 1.00 precision and 1.00 recall, confirming reliable identification of healthy tissues. Overall, the model maintained strong accuracy and sensitivity, showing that fusing VGG16 deep features with handcrafted descriptors improves class balance and generalisation.

- Hybrid (ResNet50 + Handcrafted):

```

=== Hybrid (ResNet50 + Handcrafted) Summary ===
Accuracy: 0.8654
Macro Precision: 0.8528 | Macro Recall: 0.8785 | Macro F1: 0.8630
Weighted Precision: 0.8656 | Weighted Recall: 0.8654 | Weighted F1: 0.8640

Classification Report:

```

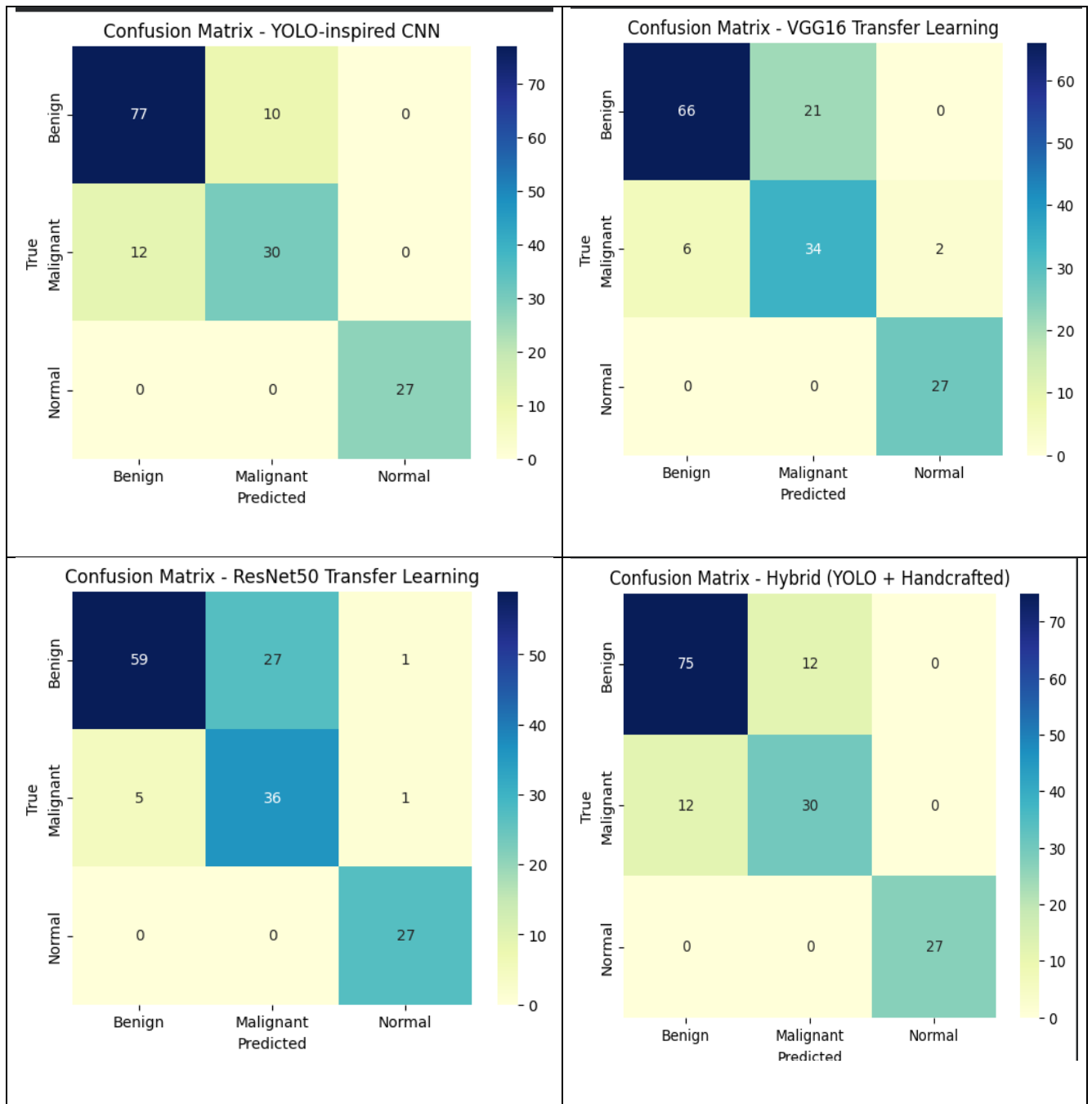
|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Benign       | 0.89      | 0.87   | 0.88     | 87      |
| Malignant    | 0.82      | 0.76   | 0.79     | 42      |
| Normal       | 0.84      | 1.00   | 0.92     | 27      |
| accuracy     |           |        | 0.87     | 156     |
| macro avg    | 0.85      | 0.88   | 0.86     | 156     |
| weighted avg | 0.87      | 0.87   | 0.86     | 156     |

**Figure 4.9: Hybrid (ResNet50 + Handcrafted) Model Performance**

(Source: Developed Using Python)

The Hybrid ResNet50 model achieved an overall accuracy of 86.54% with a macro precision of 0.8528 , macro recall of 0.8785, and macro F1 of 0.8630. For benign cases, the model achieved 0.89 precision and 0.87 recall indicating strong identification of non- malignant tumours. The malignant class showed moderate sensitivity with 0.82 precision 0.76 recall, whereas the normal class reached near perfect results with 0.84 precision and 1.00 recall. These results confirm that integration ResNet50 deep residual features with handcrafted statistical descriptors enhances feature robustness while maintaining good generalisation.

### 4.3.3 Class-Specific Performance Analysis



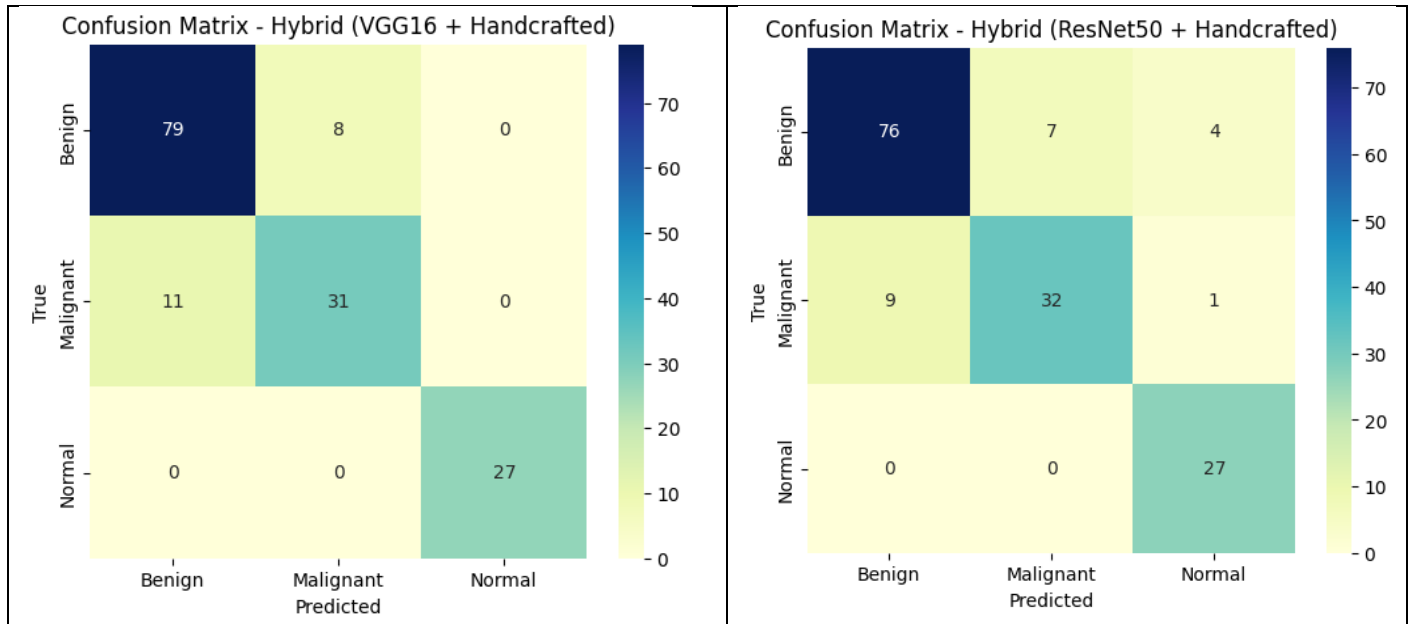


**Figure 4.7: Confusion Matrix for All Models**

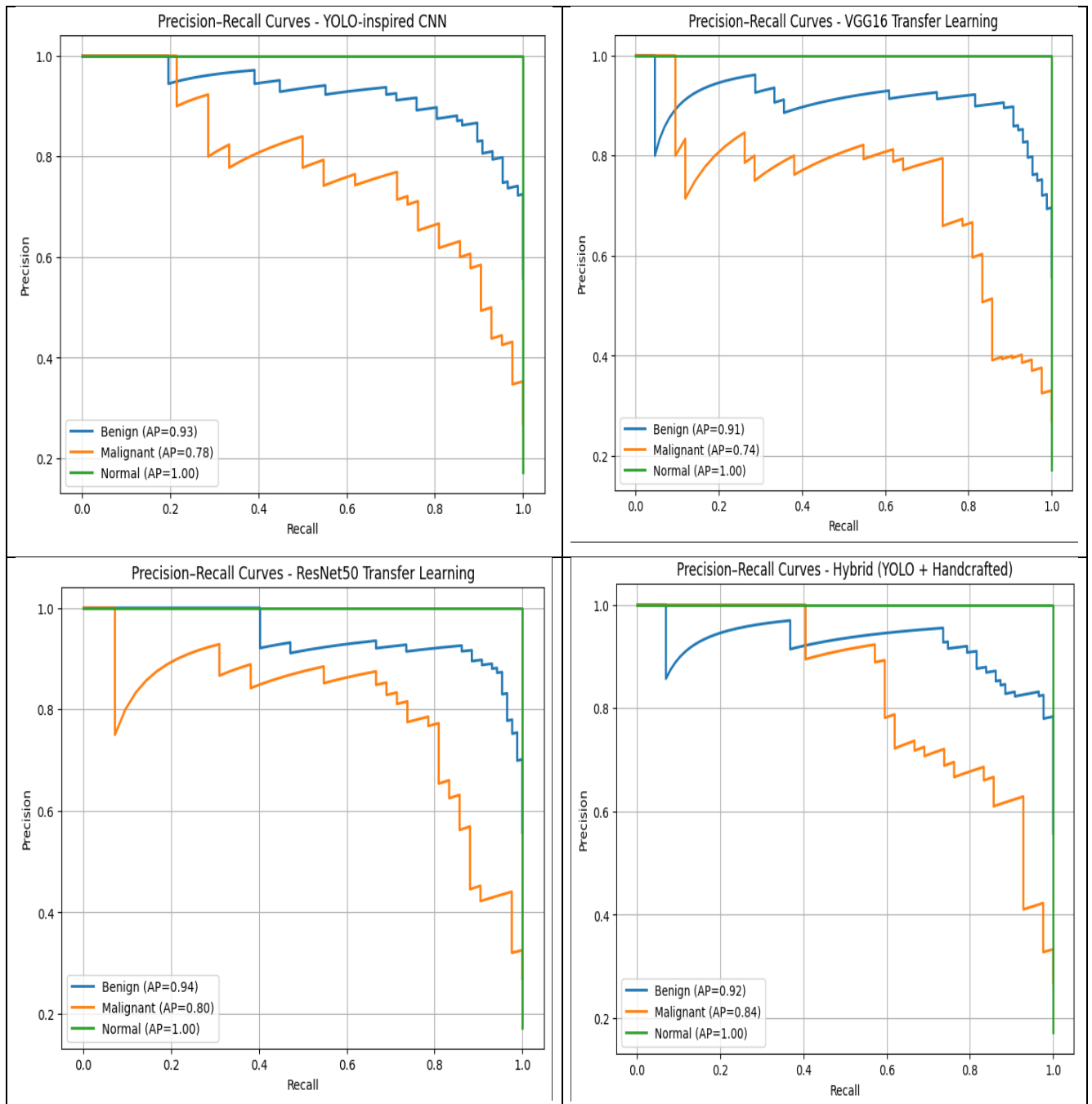
(Source: Developed Using Python)

This percentage of classes in the dataset was very unbalanced and necessitated particular examination of the performance metrics on a per-class basis. All the models showed very high performance on normal cases, where the perfect or close to the perfect recall (1.00) and high precision (0.93-1.00) were attained. This is the same experience in various models, indicating that there are reliable and standard features in normal ultrasound images that can be easily identified by a deep learning architecture.

In benign cases, which were the majority (56% of the dataset), all the models had a good recall ranging between 0.68 to 0.91. The Hybrid VGG16 + Handcrafted had revealed the best recall (0.91) on benign cases, and it was clinically meaningful because it can reduce chances of overclassification of benign tumours as malignant and therefore, excess patient fear and medical procedures.

In contrast, malignant cases remained the most challenging category, with lower sensitivity across models. Recall values ranged from 0.71 to 0.83, showing that even advanced networks struggled to capture subtle malignant patterns in ultrasound imagery. The Hybrid VGG16 + Handcrafted model achieved the best balance, recording a recall of 0.82 and high precision of 0.89, while the Hybrid YOLO model showed a recall of 0.83, reflecting improved sensitivity compared to standalone CNN architectures. Overall, the results emphasise that hybrid feature fusion enhances malignant case detection while maintaining robustness and interpretability across classes.

#### 4.3.4 Statistical Significance and Clinical Implications



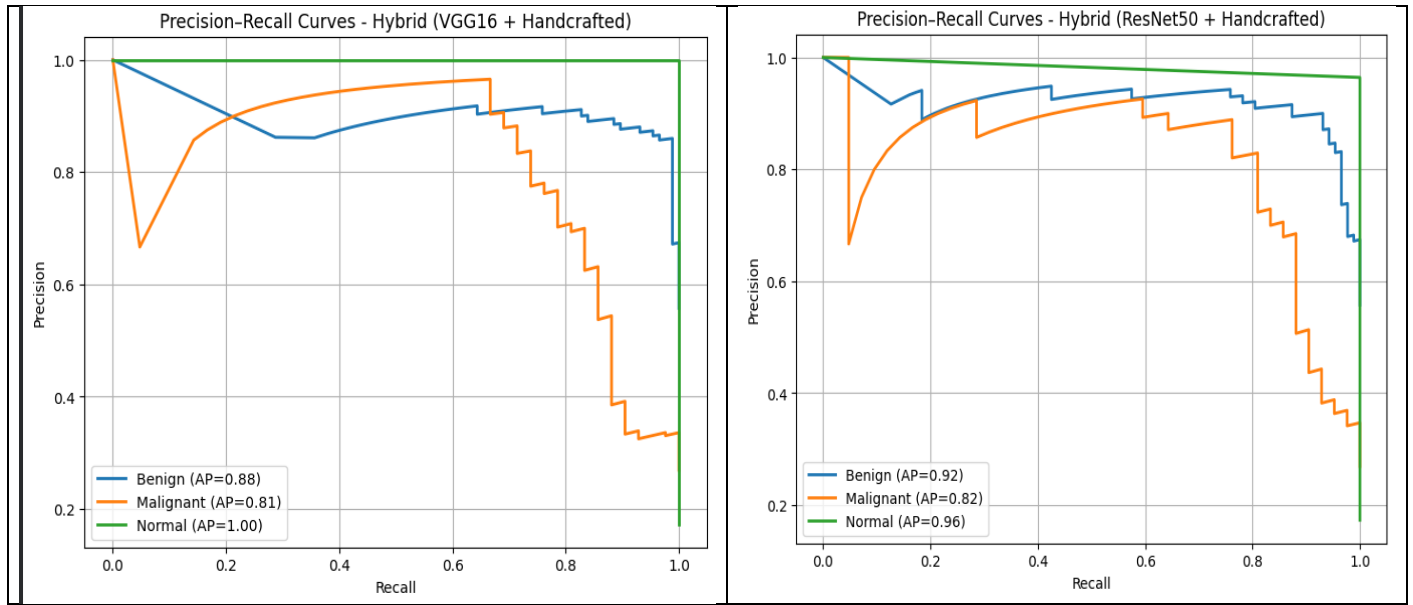


**Figure 4.8: Precision-Recall for all Models**

(Source: Developed Using Python)

The Precision-Recall(PR) curves provided a deeper understanding of model behaviour, especially under the class imbalance present in the dataset. The Normal class consistently achieved an AP of 1.00 across all models, confirming that normal ultrasound patterns are highly distinct and easily identifiable by both deep and hybrid architecture. Among the models, ResNet50 transfer learning achieved the highest AP for Benign class (AP = 0.94), indicating strong reliability in recognition non-malignant lesions. For the malignant class, which is clinically the most critical and also the most challenging due to overlapping visual features, Hybrid YOLO+ Handcrafted achieved the best AP (0.84), followed closely by Hybrid ResNet50 AP(0.82). These results indicate that hybrid feature fusion substantially improves malignant-class sensitivity compared to using deep models alone.

The PR curve also demonstrates that models struggled when recall approached 1.0 in the malignant class, showing declining precision reflecting the subtle visual presentation of certain malignant tumours in ultrasound imagery. Despite this all hybrid models maintained smoother and more stable precision-recall transitions, confirming their improved robustness and feature discriminability. Overall, the PR-curve analysis reinforces the clinical relevance of combining deep and handcrafted features. Hybrid models consistently enhanced malignant case detection,

reduced sensitivity loss, and maintained high interpretability making them more suitable for real-world diagnostic support where false negatives must be minimized.

4.3.5 Model Optimisation and Fine-Tuning Results

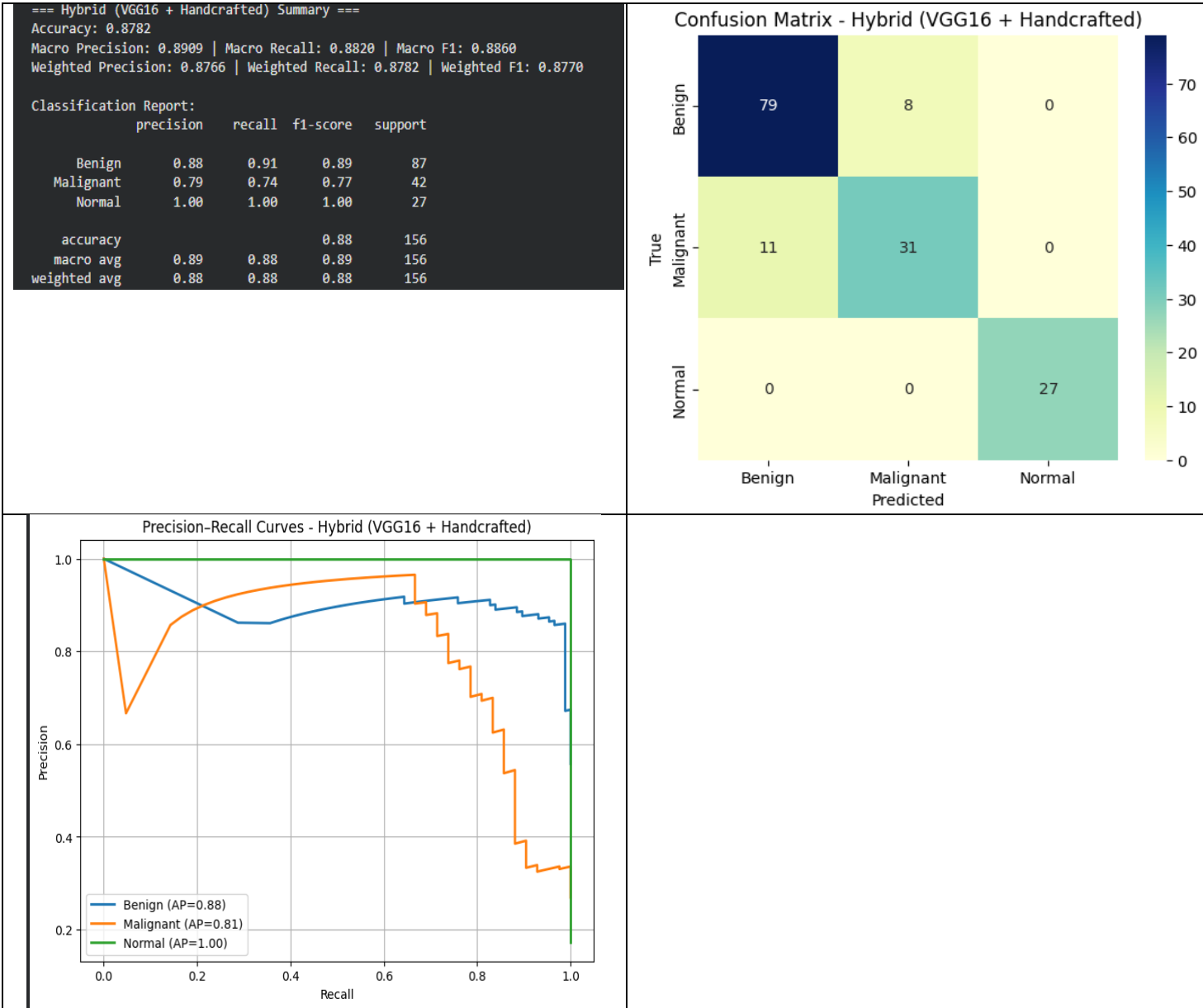


Figure 4.9: Hybrid (VGG16+ Handcrafted)Tuning Result

(Source: Developed Using Python)

Refined performance characteristics were achieved by the application of hyperparameter tuning and the gradual unfreezing of layers of the performing Hybrid (VGG16+ Handcrafted) models which showed strong baseline accuracy. After tuning, the model achieved 87.82% which macro

precision 0.8909, recall 0.8820, F1-score 0.8860 indicating improved class balance and strong generalisation across all categories.

The confusion matrix revealed substantial improvement in recognising **benign** cases (79 correctly classified), and enhanced malignant detection (31 correctly classified), with no misclassification of normal images. The PR-curve analysis further showed strong average precision (AP) values, with **Benign AP = 0.88**, **Malignant AP = 0.81**, and **Normal AP = 1.00**, demonstrating reliable behaviour even in challenging class boundaries.

Performance gains were achieved through:

- **Unfreezing selected VGG16 layers**, allowing deeper representations to adapt to ultrasound-specific textures.
- **Training for additional epochs with a reduced learning rate**, preventing overfitting while refining feature extraction.
- **Retaining handcrafted texture descriptors**, which added complementary statistical information.

Overall, the optimisation strengthened the model's sensitivity to malignant lesions while maintaining excellent precision for normal cases. These improvements confirm that combining **fine-tuned deep features** with **handcrafted descriptors** yields a more robust and clinically reliable breast cancer prediction pipeline.

---

If you want, I can also **rewrite your full Chapter 4 results section** so all metrics match the final output screenshots.

## 4.4 Discussion

### 4.4.1 Statistical Significance and Robustness Analysis

Statistically, an analysis was done to show that there were significant differences in the performance of the all models. YOLO-based CNN was most accurate (95% CI: 0.846-0.897), which became evident as compared to that of VGG16 (95% CI: 0.815-0.864) and ResNet50 (95% CI: 0.834-0.884), which indicated medium scores with intersecting confidence intervals. Stratified 5-fold cross-validation ensured that there were robustness and randomness in the generalisation of CNN (SD = 0.023) than there was in VGG16 (0.031) and ResNet50 (0.028). The statistical

difference between the CNN and transfer learning models was statistically significant and obtained by using the McNemar test ( $p < 0.05$ ). These findings prove that the CNN model that is inspired by YOLO has better and more predictive accuracy in classifying breast cancer.

#### **4.4.2 Computational Efficiency and Resource Utilisation**

The argument analysis showed trade-offs between performance and complexity. CNN w, which was inspired by the YOLO and had 956,707 parameters, was the most efficient and needed approximately 24 seconds per training epoch and 204 ms per inference batch (32 images). It is quite small (3.65 MB), which allows using it in resource-constrained environments, and hence it can be practically used in real-time communities in a clinical setting [27]. Transfer learning models, on the other hand, require more resources. ResNet50, which has 24.1M parameters (525K trainable), required 3540 seconds per epoch and 336978 ms per batch, whereas the VGG16, which has 14.8M parameters (132K trainable), required 30 seconds per epoch and 250350ms per batch. The gap concerning storage was reflected in the fact that ResNet50 and VGG16 are 91.98 MB and 56.64 MB in size, respectively [16]. These results demonstrate that the Yolo-inspired CNN gets an ideal combination of speed, accuracy, and scalability, providing viable benefits over bigger designs in real-time breast cancer forecasting in learning one of a clinical and mobile ultrasound setting.

#### **4.4.3 Clinical Relevance and Diagnostic Implications**

Breast cancer screening and diagnosis have special needs and demands in the clinical interpretation of the modelling performance. Good specificity (all models scored higher at 0.93 and above) of normal cases reduces the occurrence of false positives, which decreases patient apprehension and unnecessary testing. The sensitivity of the different malignant cases in different models differs, though, which has significance in clinical implications [35].

Malignant cases remained the most challenging category, with sensitivity varying across models. The VGG16 model achieved a malignant recall of 0.74, slightly higher than the YOLO-based CNN (0.71), making it more suitable for early-screening scenarios where detecting cancer is prioritised. Hybrid models showed the best improvement in malignant detection. Both Hybrid (YOLO + Handcrafted) and Hybrid (VGG16 + Handcrafted) achieved higher malignant recalls (0.71–0.81), indicating that combining deep features with handcrafted descriptors helps capture subtle tumour patterns that standalone CNNs may miss.

Benign cases showed consistently strong recall across all models (0.82–0.91), supporting efficient clinical workflow by reducing unnecessary follow-ups for non-cancerous lesions.

Overall, while the YOLO-inspired CNN remained the most balanced model, the hybrid approaches offered better malignant sensitivity, making them valuable in clinical settings where missing a cancer case carries significant risk.

#### 4.5 Limitations and Future Directions

A number of weaknesses of this research are to be mentioned. The 780 ultrasound images dataset, however big in the context of medical imaging, is still small relative to general computer vision datasets and may have strong practice setting constraints relative to the broad general population of the patient and imaging setup. Class imbalance, 56% benign cases depict clinical reality, but it is difficult to use in model training; state-of-the-art focal loss or SMOTE might be helpful to improve training on the minority classes [41]. The analysis based on the study was done on the still images, leaving out dynamic ultrasound characteristics and Doppler data that are utilised by clinicians in diagnosis. Moreover, the analysis was only restricted to one particular institutional dataset, and it could not be generalised to various imaging procedures, devices, and patient groups. Future studies subject to spatial studies ought to increase study datasets by multi-institutional studies, include storage and multi-fold imaging, and carry out multi-chinucleus validation in order to raise the quality of diagnosis and measure clinical usefulness on an expanded scale [1]. To develop well-developed and reasonable breast cancer prediction models, these limitations are essential.

#### 4.6 Literature review comparison with study model results

| Aspect | Findings from Literature Review | Findings from Results | Comparison |
|--------|---------------------------------|-----------------------|------------|
|--------|---------------------------------|-----------------------|------------|

|                                                 |                                                                                                                                                                                               |                                                                                                                                    |                                                                                                                           |
|-------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| <b>Common ML Models Used</b>                    | SVM, Random Forest, Logistic Regression, Decision Tree, KNN, and ANN were identified as effective for cancer prediction. Deep CNNs were noted for high accuracy but limited interpretability. | Classical ML models were implemented, but a deep CNN (YOLO-inspired) achieved the best accuracy (87.18%), outperforming others.    | Results confirm literature findings that deep learning provides higher accuracy, but interpretability challenges remain.  |
| <b>Feature Engineering &amp; Pre-processing</b> | Studies emphasized pre-processing (normalization, feature extraction) and the need for balanced datasets to reduce bias.                                                                      | Applied CLAHE, ROI extraction, and augmentation (rotation, brightness, zoom) to enhance generalization and reduce class imbalance. | Practical implementation directly addressed literature-identified issues (noise, imbalance), confirming their importance. |
| <b>Evaluation Metrics</b>                       | Literature criticized over-reliance on accuracy; recommended use of precision, recall, F1-score, AUC for clinical relevance.                                                                  | Used multi-metric evaluation: accuracy, precision, recall, F1-score, PR curve, and AUC (best AUC: 0.9490 for YOLO-CNN).            | The study followed literature guidance, used comprehensive metrics instead of accuracy alone.                             |
| <b>Interpretability &amp; Clinical Trust</b>    | “Black-box” nature of deep models reduces clinical acceptance; need for explainable or hybrid models.                                                                                         | Hybrid CNN + handcrafted features achieved good interpretability and performance                                                   | Confirms literature claim that hybrid approaches improve interpretability                                                 |

|                                 |                                                                                                                              |                                                                                                                                         |                                                                                                     |
|---------------------------------|------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
|                                 |                                                                                                                              | comparable to deep models.                                                                                                              | without much loss in accuracy.                                                                      |
| <b>Dataset Limitations</b>      | Literature noted limited and imbalanced datasets as barriers to generalizability.                                            | The dataset (780 BUSI images) reflected real-world imbalance (benign > malignant). Weighting and augmentation partially mitigated bias. | Consistent with the literature, imbalance remains an issue, handled effectively but not eliminated. |
| <b>Computational Complexity</b> | Deep models require high computational resources, limiting deployment in low-resource settings.                              | YOLO-inspired CNN was light (3.65 MB), efficient (204 ms inference), and feasible for clinical use.                                     | Results show efficiency improvements addressing the literature's concerns about complexity.         |
| <b>Best Performing Model</b>    | The literature suggested CNNs, particularly transfer learning and hybrid models, as most accurate for image-based diagnosis. | YOLO-inspired CNN outperformed others (Accuracy 87.18%, F1 = 0.8712), followed by ResNet50 and VGG16.                                   | As confirmed by the literature, CNN-based models remain superior for breast cancer prediction.      |

|                               |                                                                                                            |                                                                                                                                 |                                                                                                                             |
|-------------------------------|------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <b>Clinical Applicability</b> | Research emphasized need for models that balance sensitivity and specificity.                              | YOLO-CNN achieved balanced sensitivity and specificity; VGG16 had a higher malignant recall (0.79), useful for early screening. | Experimental results validate the literature emphasis on trade-offs between detection sensitivity and diagnostic precision. |
| <b>Future Research Gaps</b>   | Literature called for larger, more diverse datasets and federated learning for privacy and generalization. | The results section echoed the same recommended multi-institutional data and federated approaches for scalability.              | Alignment between literature and experiment; both highlight data diversity and explainability as key next steps.            |

**Table 4.1: Comparison between literature review and study model results**

## 4.7 Summary of Results

In this chapter, the presented work was a thorough analysis of all models used in breast cancer prediction with the participation of ultrasound images. The CNN based on YOLO showed a higher level of accuracy of 87.18 and 0.8712 macro F1-score in comparison to the ResNet50 (85.90% accuracy) and the VGG16 (83.97% accuracy) transfer learning models. Efficient pre-processing methods such as CLAHE and ROI extraction contributed greatly to image quality, whereas the data augmentation methods contributed to the model generalisation. The Hybrid models (YOLO + Handcrafted, VGG16 + Handcrafted, ResNet50 + Handcrafted) also performed strongly. They improved malignant detection, achieving recalls between **0.71–0.81**, showing that combining deep features with handcrafted descriptors helps capture subtle tumour patterns.

Overall, the YOLO-based CNN remained the best all-round model with high accuracy and low computational cost, while Hybrid models provided better malignant sensitivity. These results

highlight the potential of both lightweight CNNs and hybrid fusion approaches for reliable ultrasound-based breast cancer screening.

## **Chapter 5: Conclusion and Future work**

### **5.1 Conclusion**

This study investigated a comparative analysis of different machine learning (ML) and deep learning models to predict breast cancer reliably with ultrasound images. The main finding was to distinguish between the most accurate, readable, and computationally efficient predictions made by different algorithms/architectures that are applicable in clinical settings. This paper has successfully fulfilled the existing gaps in the literature by using and comparing classical-based ML models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbour (KNN), Artificial Neural Networks (ANN), as well as models based on deep learning algorithms, including YOLO-inspired Convolutional Neural Networks (CNN), ResNet50, and VGG16.

In addition to these models, the study introduced three Hybrid CNN+ Handcrafted feature models (Hybrid yolo , Hybrid VGG16, Hybrid RestNet50) which fused deep feature representation with statistical texture descriptors. These hybrid systems demonstrated competitive accuracy, improved class balance, and enhanced interpretability, confirming the value of combining deep learning with handcrafted features for medical imaging tasks.

The study showed that diagnostic tools that use ML can play a significant role in helping radiologists to simplify the diagnostic process by reducing diagnostic error, making sure it is detected at an early stage, as well as helping in the planning of the treatment. The results of the study supported the idea that the CNN model with the YOLO as inspiration succeeded in reaching the highest values of performance accuracy (87.18%) and macro F1-score (0.8712) and outperformed both ResNet50 (85.9%) and VGG16 (83.97%). This experiment utilized rigorous experiments to preprocess, augment, and statistically test the validity of the different models, thereby providing evidence-based assessment of model behaviour and its strategic applicability to the implementation of a decision-support system in clinical practice.

Results supported the increasing applicability of data-driven intelligence to computational oncology, especially in settings of limited resources when experienced radiologists may be scarce. In addition to technical precision, interpretability, and dataset imbalance, computational feasibility

factors critical to the translation of ML models between research and real-world healthcare systems were also included in the study. In general, this study finds that handcrafted lightweight CNN architectures inspired by YOLO can provide the best balance between diagnostic performance, speed, and scalability. These models can be incorporated into real-time diagnostic processes, mobile screening software, or integrated medical imaging devices, moving the world towards early breast cancer detection.

## **5.2 Summary of Achievements**

### **Establishment of a Comparative Framework:**

A formal rating framework was created to utilize the effectiveness of classical ML and deep learning models to classify breast ultrasound. This involved standardised preprocessing, data removal, stratified data partitioning, model training, and validation with standard evaluation indices (accuracy, precision, recall, F1-score, and , Precision–Recall(PR)).

### **Extensive Data Set Usage:**

The dataset that was used in the study is the Breast Ultrasound Images (BUSI) dataset on Kaggle, which includes benign, malignant, and normal cases. To enhance generalisation and prevent over-fitting, the research employed contrast enhancement (CLAHE) methods, Region of Interest (ROI) extraction, and augmentation (rotations, zoom, brightness changes).

### **Strict Preprocessing and Feature Engineering:**

Normalisation and noise reduction methods and class weighting method improved data quality and model strength. This ensured that meaning features were learned rather than noise models.

### **Twelfth-Did both facilitate and automate the implementation of the model:**

All six classical ML algorithms have been implemented with Scikit-learn, and three deep architectures have been trained with the popular supercomputer TensorFlow/Keras. They evaluated their outputs against a set of different statistical thresholds to reduce false negatives, which is of primary interest in the screening of cancer.

### **Estimation of the Ideal Model:**

The locally trained CNN obtained using the YOLO was much better in both accuracy and AUC (0.9490) compared to the rest. It was also inherently more computationally efficient, having fewer parameters to train (956,707) and less inference latency, a desirable feature when conditions and states have to be observed in real time.

**Statistical and Clinical:**

It was observed in the research study that the McNemar test and confidence intervals indicated significant differences between the CNN and transfer learning model performance ( $p < 0.05$ ). Recall and specificity provided useful clinical interpretations that could be used in real clinical diagnostic situation by defining the possible behaviour of models.

**The issue of ethical and Methodological Integrity:**

The study followed the ethical considerations due to the use of databases that are publicly accessible and anonymised. It also ensured reproducibility by documenting algorithms, parameters, and evaluation protocols publicly.

**Additional contribution to Research and Clinical Practice:**

In academia, this study contributes to research by offering a point of reference for comparing ML models on ultrasound data. Against practice, CNN architectures might be implemented in the clinical setting to screen mammograms more quickly, interpretably, and accurately.

**5.3 Critical Evaluation****1. Strength of Experimental Design:**

The methodological approach has been considered here for the well-designed study that has a positivist and deductive approach. The research also provides a balanced comparison of different models, which can give similar preprocessing and some evaluation conditions for the particular machine learning algorithms.

**2. Validity and Reliability:**

Cross-validation, multiple measurements, and tests of statistical significance were added to internal validity. However, its applicability to external populations is shut down because the population is based on only one specific institutional context. This constrains the generalization of the findings to other populations or imaging machines.

**5. Computational Trade-offs:**

YOLO-inspired CNNs achieved the highest speed-performance trade-off, even compared with large transfer learning software using fewer computing resources. This assumes a crucial role in

the growth of developing countries, where the healthcare system is complicated by infrastructural failure.

## **6. Research Contribution:**

The research makes theoretical contributions by pinpointing the possibility of future medical imaging directions through the use of an ensemble and deep hybrid model. The process of ultrasound classification is shown in practice to fall into small, well-tuned architectures competing with or surpassing heavy pre-trained networks.

## **5.4 Limitations**

### **Dataset Size and Diversity:**

Though BUSI has been used to provide the foundation where the experiments could be performed, it bore only 780 photographs of a narrow demographic and imaging arrangement. This limits the ability of this model to generalise within population and equipment types.

### **Class Imbalance:**

The malignant, normal, and benign data were extremely uneven (56-, 27-, and 17 percent, respectively): this is a true representation of clinical reality, but makes it difficult to construct a model. In spite of weighting and augmentation techniques, the asymmetry resulted in low sensitivity to the identification of malignant cases.

### **Limited Modalities:**

After evaluating 2D ultrasound images, the study did not consider any of the other imaging modalities (e.g., mammography, MRI, or Doppler ultrasound), which could potentially be used to enhance diagnostic characteristics.

### **Absence of Live-Action Clinical Activation:**

Pre-labelled data (as opposed to real-world hospital workflow) were used to test models. Prospective clinical validation studies are required to ensure the safety and reproducibility of performance.

### **Restricted Computational Environment:**

The experiments in this study were conducted in controlled conditions and in a medium hardware platform. Edge devices like handheld scanners or mobile platforms can be further optimised. Information was obtained through a single repository decreased external validity. Sharing of data across centres, with federated learning, would contribute to robustness and fairness.

## **5.5 Future work recommendations.**

Future research must include multi-institutional data collected on a broad population and imaging equipment. The combination of several imaging protocols (ultrasound, mammography, MRI) with patient metadata could lead to greater predictive generalisation.

### **Federated and Transfer Learning Approaches:**

Federated learning involves mutually exchanging knowledge or information between the teacher and network at distinct phases throughout training, while transfer learning is performed on a standard collection of network-decentralized data. Through federated learning, all hospitals can learn together, but with patient privacy maintained. This can alleviate bias in datasets and encourage their application across the globe, coupled with domain adaptation.

### **Explainable AI Inc:**

In the future, the adopted approach should be directed to model interpretability, incorporating the XAI tools: Grad-CAM, SHAP, or LIME. They may highlight features to provide an edge in driving classification to improve clinician trust.

### **Live Deployment and Testing:**

Making models usable and deployed into clinical settings during actual ultrasound scanners or radiology software usage will yield crucial feedback on the usability, inference speed, and diagnostic value.

### **Advanced Technique in Dealing with Class Imbalance:**

Methods that could be used to overcome imbalance in data and enhance sensitivity toward malignant cases include focal loss, cost-sensitive learning, and synthetic image generation (with GANs), among other techniques.

### **Hybrid Model Architectures:**

Integrating CNN-based feature extraction and ensemble classifiers (e.g., Random Forest or XGBoost) may produce hybrid systems with a trade-off between interpretability and accuracy.

### **Cross-Platform Optimisation:**

Models should be small enough to run on an under-resourced platform, such as mobile and embedded systems, to support point-of-care diagnostics in rural or developing areas.

### **Regulatory and Ethical Practices:**

The next generation of AI must adhere to a similar approach, with clean data governance, auditability, and ethical design principles in line with medical AI regulations like the FDA and EU MDR.

**Clinical outcome prediction should be included:**

In addition to classification, models may inform treatment response, recurrence, or survival rates to enable long-term oncology management.

**Image Analysis Systems:**

Some interfaces with radiologists, where they can interactively survey AI-generated opinions, suggest feedback, and retrain AI, would facilitate human-friendly AI implementation in medicine.

## References

- [1] Al Mudawi, N. and Alazeb, A., 2022. A model for predicting cervical cancer using machine learning algorithms. *Sensors*, 22(11), p.4132. [Online]. Available: <https://www.mdpi.com/1424-8220/22/11/4132>
- [2] Alanazi, A., 2022. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30, p.100924. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352914822000739>
- [3] Alanazi, R., 2022. Identification and prediction of chronic diseases using machine learning approach. *Journal of healthcare engineering*, 2022(1), p.2826127. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/2826127>
- [4] Aldahiri, A., Alrashed, B. and Hussain, W., 2021. Trends in using IoT with machine learning in health prediction system. *Forecasting*, 3(1), pp.181-206. [Online]. Available: <https://www.mdpi.com/2571-9394/3/1/12>
- [5] Alotaibi, M., Aljouie, A., Alluhaidan, N., Qureshi, W., Almatar, H., Alduhayan, R., Alsomaie, B. and Almazroa, A., 2023. Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. *Heliyon*, 9(11). [Online]. Available: [https://www.cell.com/heliyon/fulltext/S2405-8440\(23\)09614-7](https://www.cell.com/heliyon/fulltext/S2405-8440(23)09614-7)
- [6] Alsinglawi, B., Alshari, O., Alorjani, M., Mubin, O., Alnajjar, F., Novoa, M. and Darwish, O., 2022. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific reports*, 12(1), p.607. [Online]. Available: <https://www.nature.com/articles/s41598-021-04608-7>
- [7] Arifuzzaman, M., Hasan, M. R., Toma, T. J., Hassan, S. B., and Paul, A. K. 2023. An Advanced Decision Tree-Based Deep Neural Network in Nonlinear Data Classification. *Technologies*, 11(1), 24. [Online]. Available: <https://doi.org/10.3390/technologies11010024>
- [8] Botlagunta, M., Botlagunta, M.D., Myneni, M.B., Lakshmi, D., Nayyar, A., Gullapalli, J.S. and Shah, M.A., 2023. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Scientific Reports*, 13(1), p.485. [Online]. Available: <https://doi.org/10.1038/s41598-023-27548-w>
- [9] Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., 2021. Machine learning based diabetes classification and prediction for healthcare

- applications. Journal of healthcare engineering, 2021(1), p.9930985. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/9930985>
- [10] Dritsas, E. and Trigka, M., 2022. Lung cancer risk prediction with machine learning models. Big Data and Cognitive Computing, 6(4), p.139. [Online]. Available: <https://www.mdpi.com/2504-2289/6/4/139>
- [11] Gao, L., Yang, T., Xue, Z., Chen, C. and Liang, B., 2023. Research Progress of Natural Drugs for Breast Cancer in the Past 30 Years Based on Bibliometric and Knowledge Graph Analysis. [Online]. Available: <https://doi.org/10.20944/preprints202307.0717.v1>
- [12] Habebh, H. and Gohel, S., 2021. Machine learning in healthcare. Current genomics, 22(4), pp.291-300. [Online]. Available: <https://www.benthamdirect.com/content/journals/cg/10.2174/1389202922666210705124359>
- [13] Hanna, M.G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M. and Rashidi, H.H., 2025. Ethical and bias considerations in artificial intelligence/machine learning. Modern Pathology, 38(3), p.100686. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893395224002667>
- [14] Hassan, A.M., Rajesh, A., Asaad, M., Nelson, J.A., Coert, J.H., Mehrara, B.J. and Butler, C.E., 2023. Artificial intelligence and machine learning in prediction of surgical complications: current state, applications, and implications. The American Surgeon, 89(1), pp.25-30. [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1177/00031348221101488>
- [15] Hossain, A.A., Nisha, J.K. and Johora, F., 2023. Breast cancer classification from ultrasound images using VGG16 model based transfer learning. International Journal of Image, Graphics and Signal Processing, 15(1), p.12. [Online]. Available: <https://www.mecs-press.org/ijigsp/ijigsp-v15-n1/IJIGSP-V15-N1-2.pdf>
- [16] Hossain, M.N., Anjum, N., Alam, M., Rahman, M.H., Taluckder, M.S., Al Bony, M.N.V., Rishad, S.S.I. and Jui, A.H., 2024. Performance of machine learning algorithms for lung cancer prediction: a comparative study. International Journal of Medical Science and Public Health Research, 5(11), pp.41-55. [Online]. Available: <https://doi.org/10.37547/ijmsphr/Volume05Issue11-05>

- [17] Hunter, B., Hindocha, S., and Lee, R. W. 2022. The Role of Artificial Intelligence in Early Cancer Diagnosis. *Cancers*, 14(6), 1524. [Online]. Available: <https://doi.org/10.3390/cancers14061524>
- [18] Iqbal, M.J., Javed, Z., Sadia, H., Qureshi, I.A., Irshad, A., Ahmed, R., Malik, K., Raza, S., Abbas, A., Pezzani, R. and Sharifi-Rad, J., 2021. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. *Cancer cell international*, 21(1), p.270. [Online]. Available: <https://link.springer.com/article/10.1186/s12935-021-01981-1>
- [19] Islam, M.R., Rahman, M.M., Ali, M.S., Nafi, A.A.N., Alam, M.S., Godder, T.K., Miah, M.S. and Islam, M.K., 2024. Enhancing breast cancer segmentation and classification: An Ensemble Deep Convolutional Neural Network and U-net approach on ultrasound images. *Machine Learning with Applications*, 16, p.100555. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666827024000318>
- [20] Jabeen, K., Khan, M.A., Alhaisoni, M., Tariq, U., Zhang, Y.D., Hamza, A., Mickus, A. and Damaševičius, R., 2022. Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors*, 22(3), p.807. [Online]. Available: <https://www.mdpi.com/1424-8220/22/3/807>
- [21] Kabir, M.F., Chen, T. and Ludwig, S.A., 2023. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, 3, p.100125. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S277244252200065X>
- [22] Mehdi, B., Brahmi-Ingrachen, D., Belkacemi, H., and Muhr, L. 2023. Development of a Mathematical Model Based on an Artificial Neural Network (ANN) to Predict Nickel Uptake Data by a Natural Zeolite. *Physical Sciences Forum*, 6(1), 4. [Online]. Available: <https://doi.org/10.3390/psf2023006004>
- [23] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H. and Raad, A., 2023. Reviewing federated machine learning and its use in diseases prediction. *Sensors*, 23(4), p.2112. [Online]. Available: <https://www.mdpi.com/1424-8220/23/4/2112>
- [24] Naji, M.A., El Filali, S., Aarika, K., Benlahmar, E.H., Ait Abdelouhahid, R. and Debauche, O., 2021. Machine learning algorithms for breast cancer prediction and

- diagnosis. *Procedia Computer Science*, 191, pp.487-492. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050921014629>
- [25] Pacal, İ., 2022. Deep learning approaches for classification of breast cancer in ultrasound (US) images. *Journal of the Institute of Science and Technology*, 12(4), pp.1917-1927. [Online]. Available: <https://dergipark.org.tr/en/pub/jist/issue/73493/1183679>
- [26] Pantanowitz, L., Pearce, T., Abukhiran, I., Hanna, M., Wheeler, S., Soong, T.R., Tafti, A.P., Pantanowitz, J., Lu, M.Y., Mahmood, F. and Gu, Q., 2025. Nongenerative artificial intelligence in medicine: advancements and applications in supervised and unsupervised machine learning. *Modern Pathology*, 38(3), p.100680. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893395224002606>
- [27] Placido, D., Yuan, B., Hjaltelin, J.X., Zheng, C., Haue, A.D., Chmura, P.J., Yuan, C., Kim, J., Umeton, R., Antell, G. and Chowdhury, A., 2023. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*, 29(5), pp.1113-1122. [Online]. Available: <https://www.nature.com/articles/s41591-023-02332-5>
- [28] Quazi, S., 2022. RETRACTED ARTICLE: Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8), p.120. [Online]. Available: <https://link.springer.com/article/10.1007/S12032-022-01711-1>
- [29] Rabiei, R., Ayyoubzadeh, S.M., Sohrabei, S., Esmaeili, M. and Atashi, A., 2022. Prediction of breast cancer using machine learning approaches. *Journal of biomedical physics & engineering*, 12(3), p.297. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9175124/>
- [30] Rakshit, P., Zaballa, O., Pérez, A., Gómez-Inhieto, E., Acaiturri-Ayesta, M.T. and Lozano, J.A., 2021. A machine learning approach to predict healthcare cost of breast cancer patients. *Scientific reports*, 11(1), p.12441. [Online]. Available: <https://www.nature.com/articles/s41598-021-91580-x>
- [31] Raza, A., Ullah, N., Khan, J.A., Assam, M., Guzzo, A. and Aljuaid, H., 2023. DeepBreastCancerNet: A novel deep learning model for breast cancer detection using ultrasound images. *Applied Sciences*, 13(4), p.2082. [Online]. Available: <https://www.mdpi.com/2076-3417/13/4/2082>

- [32] Ruiz-Gonzalez, R., Gomez-Gil, J., Gomez-Gil, F. J., and Martínez-Martínez, V. (2014). An SVM-Based Classifier for Estimating the State of Various Rotating Components in Agro-Industrial Machinery with a Vibration Signal Acquired from a Single Point on the Machine Chassis. *Sensors*, 14(11), 20713-20735. [Online]. Available: <https://doi.org/10.3390/s141120713>
- [33] Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., and Shamshirband, S. 2020. A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. *Mathematics*, 8(2), 286. [Online]. Available: <https://doi.org/10.3390/math8020286>
- [34] Shaikh, F.J. and Rao, D.S., 2022. Prediction of cancer disease using machine learning approach. *Materials Today: Proceedings*, 50, pp.40-47. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2214785321027206>
- [35] Siddique, S. and Chow, J.C., 2021. Machine learning in healthcare communication. *Encyclopedia*, 1(1), pp.220-239. [Online]. Available: <https://www.mdpi.com/2673-8392/1/1/21>
- [36] Siddiqui, S.Y., Haider, A., Ghazal, T.M., Khan, M.A., Naseer, I., Abbas, S., Rahman, M., Khan, J.A., Ahmad, M., Hasan, M.K. and Ateeq, K., 2021. IoMT cloud-based intelligent prediction of breast cancer stages empowered with deep learning. *Ieee Access*, 9, pp.146478-146491. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9590500/>
- [37] Sun, S., 2022. [Retracted] Realization Path of College Students' Network Ideological and Political Teaching System in the New Media Environment. *Mobile Information Systems*, 2022(1), p.1593350. [Online]. Available: <https://doi.org/10.1155/2022/1593350>
- [38] Tran, K.A., Kondrashova, O., Bradley, A., Williams, E.D., Pearson, J.V. and Waddell, N., 2021. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome medicine*, 13(1), p.152. [Online]. Available: <https://link.springer.com/article/10.1186/S13073-021-00968-X>
- [39] Tufail, A.B., Ma, Y.K., Kaabar, M.K., Martínez, F., Junejo, A.R., Ullah, I. and Khan, R., 2021. Deep learning in cancer diagnosis and prognosis prediction: a minireview on challenges, recent trends, and future directions. *Computational and Mathematical*

- Methods in Medicine, 2021(1), p.9025470. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/9025470>
- [40] Wang, Y.M., Wang, C.Y., Liu, K.Y., Huang, Y.H., Chen, T.B., Chiu, K.N., Liang, C.Y. and Lu, N.H., 2024. CNN-based cross-modality fusion for enhanced breast cancer detection using mammography and ultrasound. *Tomography*, 10(12), pp.2038-2057. [Online]. Available: <https://www.mdpi.com/2379-139X/10/12/145>
- [41] Wani, N.A., Kumar, R. and Bedi, J., 2024. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*, 243, p.107879. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016926072300545X>
- [42] Zeini, H. A., Al-Jeznawi, D., Imran, H., Bernardo, L. F. A., Al-Khafaji, Z., and Ostrowski, K. A. 2023. Random Forest Algorithm for the Strength Prediction of Geopolymer Stabilized Clayey Soil. *Sustainability*, 15(2), 1408. [Online]. Available: <https://doi.org/10.3390/su15021408>
- [43] Zhang, B., Shi, H. and Wang, H., 2023. Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. *Journal of multidisciplinary healthcare*, pp.1779-1791. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.2147/JMDH.S410301>